ESD-TR-87-026

Technical Report
773

A Two-Stage Isolated-Word Recognition System Using Discriminant Analysis

E.A. Martin

5 August 1987

Lincoln Laboratory

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LEXINGTON, MASSACHUSETTS

**Prepared for the Defense Advanced Research Projects Agency
under Electronic Systems Division Contract F19628-85-C-0002.**

The work reported in this document was performed at Lincoln Lahoratory, a center for research operated by Massachusetts Institute of Technology. This work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract F19628-85-C-0002 (ARPA Order 5328).

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

A TWO-STAGE ISOLATED-WORD RECOGNITION SYSTEM USING DISCRIMINANT ANALYSIS*

E.A. MARTIN
Group 24

TECHNICAL REPORT 773

5 AUGUST 1987

*This report is based on the thesis of the same title submitted to the Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology in May 1986 in partial fulfillment for the degree of Master of Science.

LEXINGTON MASSACHUSETTS

ABSTRACT

This report describes a two-stage isolated-word recognition system using a Hidden Markov Model (HMM) recognizer in the first stage, and a statistical discriminator in the second stage. The second-stage system performs pairwise discriminations between the top few candidate word models when no clear decision is made from the first stage. Likelihood-ratio comparisons and a new technique called "sifting" are used to focus attention on those features that best differentiate word pairs.

This system alleviates four fundamental problems which are found with most conventional speech recognition systems. These problems include: (1) the effects of limited training data are not explicitly taken into account, (2) the correlation between adjacent observation frames is incorrectly modeled, (3) durations of acoustic events are poorly modeled, and (4) features which might be important in discriminating only among specific word pairs or sets of words are not easily incorporated into the system without degrading overall performance. The system was tested on a 35 word/10,000 token stressed-speech isolated-word data base created at Lincoln Laboratory. The adding of the second-stage discriminating system reduced the error rate by more than a factor of 2. The overall error rate fell from 7.7 percent with only the HMM system to 3.5 percent with both the HMM system and the discriminator.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

A TWO-STAGE ISOLATED-WORD RECOGNITION SYSTEM USING DISCRIMINANT ANALYSIS

1. INTRODUCTION

1.1 AUTOMATIC SPEECH RECOGNITION

Machine recognition of human speech has proved to be an elusive goal sought after for many years. Many approaches toward this goal have been taken. The prevailing approaches reduce the problem to some form of template or pattern matching.[1,2] The difficulty in forming templates and developing models for continuous speech is considerably greater than the difficulty encountered when dealing with just isolated words. Because of this fact, a vast majority of work in this field has addressed the problem of recognizing isolated words. Dynamic Time Warping (DTW) and Hidden Markov Modeling (HMM) are representative examples of the approaches scientists have taken toward speech recognition.

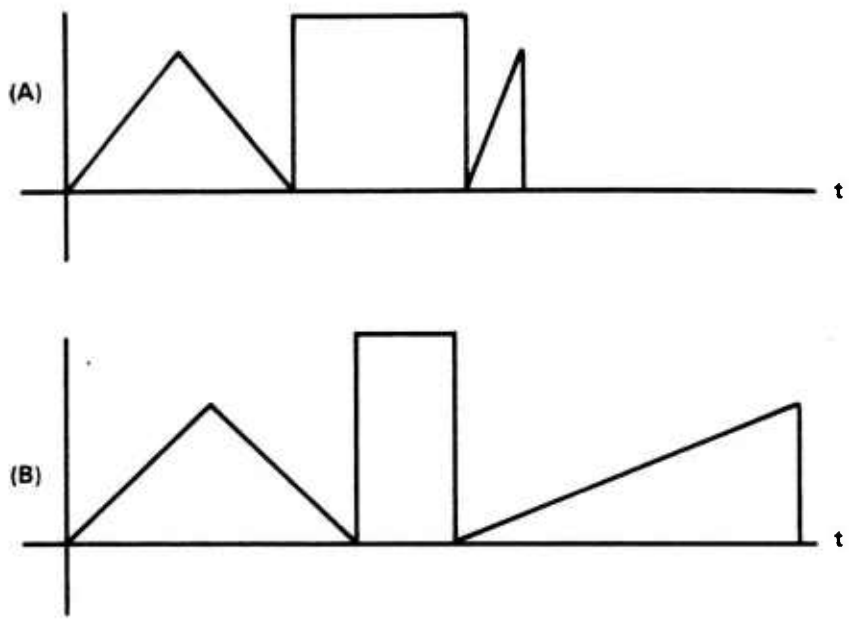
1.1.1 Dynamic Time Warping

Dynamic Time Warping is a process that time aligns an input token as it is compared to a reference word template. This is done by effectively squeezing or stretching parts of an input word to best match the durations of individual parts of the reference words. Consider Figure 1-1. Waveform (A) is some input, and (B) is some reference template.

Intuitively, similarities are seen in the waveforms. However, if a comparison scheme is used where the observation of the input at time t_0 is compared directly with the reference template at time t_0, the similarities in the two waveforms will not be effectively identified.

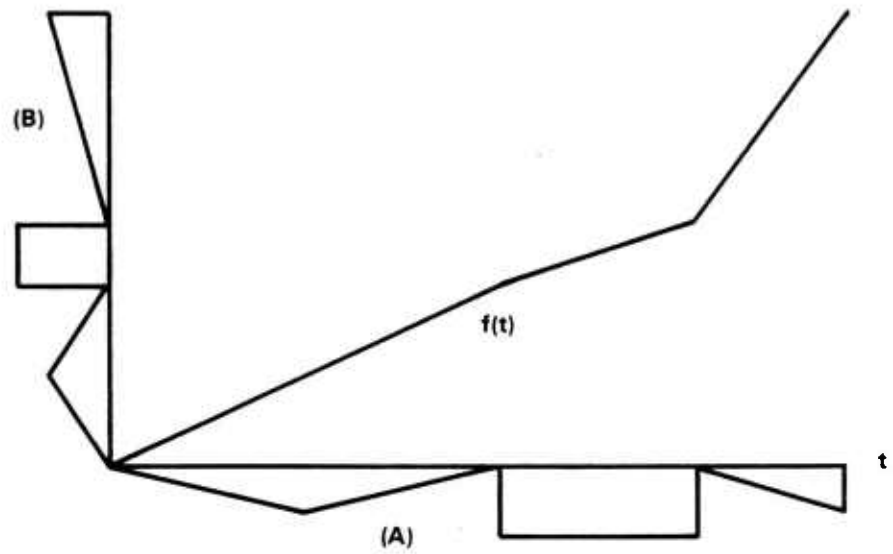
Consider another comparison scheme, as presented in Figure 1-2. If the scheme is now to compare the input at time t_0 with the reference template at time $f(t_0)$, the obvious similarities in the waveforms will be effectively identified. The matching problem then reduces to finding a warping function $f(t_0)$, given an input token and a reference template, so that the overall comparison is optimized. This will result in a much more robust pattern-matching scheme.

This describes the essential motivation for DTW. When applying DTW to speech recognition, the inputs and references are usually not observed as continuous waveforms, but rather as a discrete series of observation vectors, sampled at an appropriate frame interval. These vectors can consist of Linear Predictive Coding coefficients, cepstral coefficients, or a variety of other sets of speech parameters. Given an input observation sequence and a stored reference sequence, dynamic programming techniques are used to calculate the optimal f(t) or time warping path to optimize the comparison.



78054-14

Figure 1-1. Unaligned input and reference template.



78054-15

Figure 1-2. f(t) indicates optimal warping path between the input waveform (B) and the reference template (A).

1.1.2 Hidden Markov Modeling

Hidden Markov Modeling is another speech recognition technique that has become increasingly popular in the last few years.[3-7] HMM will be discussed in more detail in the sections to follow. The HMM technique assumes that an underlying Markov process generates the speech signal. The states of this model can generally be thought of as the acoustic events in a given word. Models are made to characterize the probability distributions of observation vectors for each state. The term "Hidden" arises from the fact that only these vectors, and not the underlying states, can be observed. These conditional probability distributions are of the form

$$\Pr(\underline{O}/a) = \Pr(\text{observing vector } \underline{O}/\text{in state } a) \quad . \tag{1.1}$$

As an input sequence comes in for some unknown word in the recognition process, each input observation vector is assigned to a node (or state in the Markov process) in a way which optimizes the alignment of input vectors with the HMM states. The overall score is computed as the product of T probability scores, where T is the number of observations in the input. Each of these scores represents the probability of a single observation vector, given its assigned state.

Given models such as this for all the vocabulary words, probability scores are computed that reflect the likelihood of a word model, given a set of observations. The recognizer then chooses the model with the highest likelihood score as the selected word.

1.2 PROBLEMS WITH RECOGNITION SYSTEMS

One key problem with DTW and HMM recognition systems is that scoring is done on a per-observation basis, where equal weighting is given to each input observation vector. Scoring of this nature assumes an underlying statistical independence of observation inputs. This is a poor assumption because observations corresponding to a vowel in steady state, for example, will be highly correlated. One technique that eliminates the assumption of statistical independence is to score on a per-HMM-node basis instead of a per-frame basis. This technique is based on the idea that each HMM node represents a unique acoustic event which is independent of adjacent events.

A second problem with many recognition systems is that the duration of acoustic events are modeled poorly. DTW time aligns the input to a reference template in a manner which suppresses the differences in durations of acoustic events. Thus, the durational characteristic is not used very effectively. Most HMM recognizers model node residency times as a decaying exponential. This implies that short acoustic events will tend to produce better scores than long acoustic events. A better model for duration as a statistic and for its use in a recognition system will be proposed, based on the actual distributions of these node residency times estimated during training.

A third problem with many recognition systems is that the effects of limited training data on estimates of recognition system model parameters are often not effectively taken into account. For example, estimates of the variances associated with the probability distribution in Equation (1.1) can be very inaccurate with small training sets. However, the means of these distributions can be estimated much more reliably. Reliance on inaccurately estimated parameters in a

recognition system can cause significant performance degradation. The use of grand variance estimates[5,8] and T-testing, to eliminate poor statistics from scoring, will be presented as a technique for solving this problem.

A fourth problem with many recognition systems is that performance tends to degrade when special-feature parameters are added to aid in discriminating a few important confusions.[9] The addition of special features in many systems degrades the scoring for most cases except the few where the feature was meant to be effective. The T-testing technique, noted above, allows features that do not contribute robustly to the scoring to be automatically omitted from the decision process. This enables special features to be included in such a way that they can resolve important confusions without creating significant additional confusions.

1.3 THE TWO-STAGE RECOGNIZER

The architecture for the new system includes two stages. The baseline or first-stage system classifies the input as one of a small subset of possible words. The second stage makes a decision by discriminating between the more limited set of candidates.

The two-stage recognition scenario is as follows. An unknown token is passed to the baseline recognizer. The recognizer performs some form of template matching between the input token and the models stored for each word in the vocabulary. Distance scores are computed, reflecting the match between the input token and each of the respective word models. Assume in this particular case the distance scores produce no clear winner because the scores for the top few word models are close. The baseline system must either reject the input token as unrecognized (not a wise decision if the distance scores are good), or just pick the model with the best score. Instead, the discriminant system will, at this point, deduce that it must choose between these close candidates, then look at those parts of the input token, both spectrally and temporally, that might better differentiate between the candidates.

A realistic example of how this system could be effective is with the "go/no" confusion. Many recognizers consistently confuse these words. A baseline system, as described above, will give strong scores for both "go" and "no," given the input is one of those words. The scores will be good since the longer part of the word (the long vowel) will match closely with both word models. Suppose the input token was "go" but the vowel section for this word matched closer with the vowel section of the "no" word model, the recognizer might give the "no" word a better score as a result of the vowel dominating the duration of the word.

The discriminant system will not weight the contribution from the long vowel significantly. This seems reasonable since it corresponds to the same phonetic event for both words. It will, however, weight the beginning of the word significantly. From training, statistics will suggest that a greater difference can be found with this part of the two candidate words. In this way the discriminant system will focus attention toward the section of the word that will best enable discrimination.

1.4 THE ROBUSTNESS ISSUE IN SPEECH RECOGNITION

One of the challenges of this system is to maintain high accuracy recognition in the presence of high acoustic noise and severe psychological and physical stress on the speaker.[10,11] To address this problem, the system presented includes the following features: (a) an improved model for the duration of acoustic events, (b) an implementation which allows confusion-specific features to be added into the recognition algorithm without degrading overall performance, and (c) a recognition algorithm which inherently focuses itself to parts of the input which most effectively allow discrimination between recognized words.

Data bases have been developed at Texas Instruments and Lincoln Laboratory to facilitate the development of recognition algorithms in this scenario. These data bases include speech from several speakers produced during a difficult workload task, when talkers are in a noise background, and when talkers use several different talking styles.

1.5 SUMMARY OF THIS REPORT

This report describes the design and implementation of such a discriminant system. Section 2 describes the baseline system that was used for all experiments. A more detailed description of HMM theory is also presented. Section 3 describes the Lincoln stressed-speech data base that was used for all experiments. It also describes the results of some preliminary testing done with this data base that indicated the potential for a two-stage discriminant system. Section 4 describes the basic scenario for using discriminant analysis with speech, the processing needed to implement such a system, and various schemes for attaining improved results from such a system. Section 5 discusses the experiments that were performed. It describes various modifications to the system presented in Section 4, and the results that were achieved from each of these experiments. Section 6 discusses further proposed experiments and the feasibility of a discriminant system as part of a large vocabulary recognizer. Conclusions and a Summary are provided in Section 7.

2. THE BASELINE SYSTEM

The system that was used as the first-stage recognizer is a maximum-likelihood-based Hidden Markov Model recognition system with continuous observations. It is described in Reference 5. To facilitate an understanding of this system, a brief description of Hidden Markov Models will be given.

In a typical Markov process a discrete symbol is observed corresponding directly to the state of that process. A state transition then occurs. These transitions are modeled by a stochastic state transition matrix that describes the probability of jumping to any state, given the previous state. Let A be the transition matrix for some model.

$$a_{ij} = \Pr(\text{state at time } t+1 \text{ is } j / \text{state at time } t \text{ is } i) \quad . \tag{2.1}$$

Given several Markov models (state transition matrices) and an observation sequence, we can find the model that produces the given observation sequence with the highest probability.

Example:

Assume an observation sequence O = A, C, D, A, B, B, A, C where A, B, C, and D are Markov states. Assume two candidate Markov models represented by their transition matrices P and Q where $P_{x,y}$ is the state transition probability of going from state x to state y under model P. The probabilities of observing the sequence, given these models, are as follows:

$$\Pr(O/\text{model P}) = (P_{a,c})\ (P_{c,d})\ (P_{d,a})\ (P_{a,b})\ (P_{b,b})\ (P_{b,a})\ (P_{a,c}) \tag{2.2}$$

$$\Pr(O/\text{model Q}) = (Q_{a,c})\ (Q_{c,d})\ (Q_{d,a})\ (Q_{a,b})\ (Q_{b,b})\ (Q_{b,a})\ (Q_{a,c}) \quad . \tag{2.3}$$

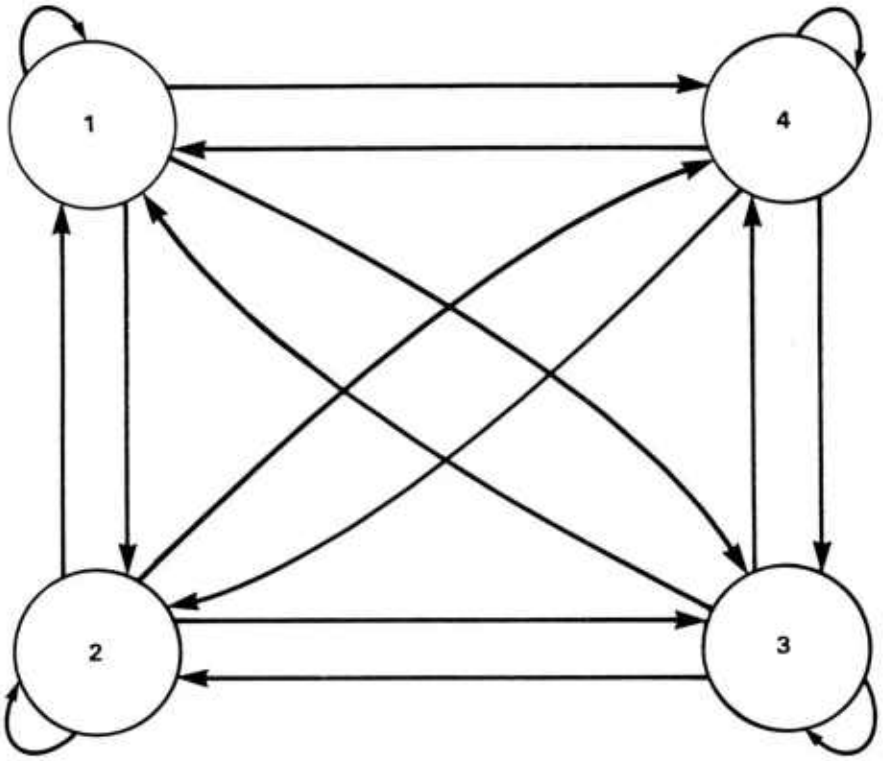
If $\Pr(O/\text{model P}) > \Pr(O/\text{model Q})$, it is said that model P is the more likely candidate because that model has a higher probability of producing the observation O than does model Q.

The key difference between Markov processes and Hidden Markov processes is that the symbol emitted from the Markov process is a deterministic function of the state the process is in; whereas, in the Hidden Markov process the symbol emitted is the result of some probabilistic function of the state.[4,5,8,12]

$$\Pr(s/i) = \Pr(\text{emitting symbol } s / \text{the state is } i) \quad . \tag{2.4}$$

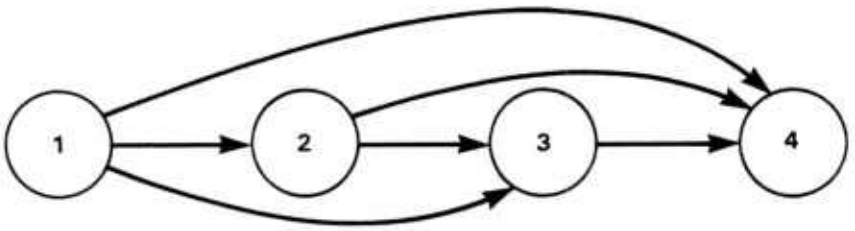
The specification of the Hidden Markov Model must include both the transition matrix (2.3) and the conditional distributions (2.4). The problem of calculating Pr(O/model) is now more complicated. However, with some assumptions as to the starting state and by some restrictions on the allowable state transition probabilities, this problem can be solved in a reasonably straightforward manner.

If the transition matrix is unrestricted, the probability of going from any state to another can be some finite probability, as illustrated in Figure 2-1. If we think of the states as being part of a sequential process where one cannot return to a previously visited state, the transition network is simplified as shown in Figure 2-2. If we further restrict the process by not allowing any



78054-11

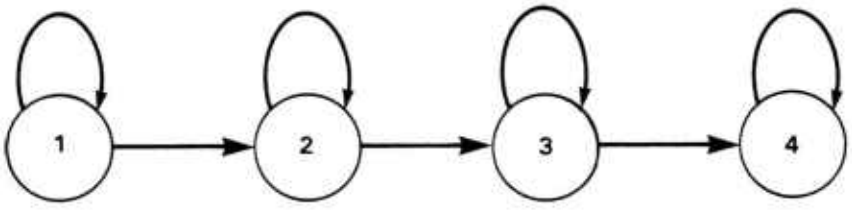
Figure 2-1. Allowable transition paths of an unconstrained Markov model.



78054-13

Figure 2-2. Allowable transition paths of a left-to-right Markov model with jump states.

state to be skipped, the network is even further simplified as in Figure 2-3. The scheme in Figure 2-3 shows the restrictions that have been used for the application of HMMs to speech in the baseline system. With this restriction the process of finding the most likely state sequence, given a model and an observation set, becomes a segmentation procedure where the number of observation frames assigned to each node are the parameters to be determined.



78054-12

Figure 2-3. Allowable transitions for a left-to-right Markov model without jump states.

In applying HMM techniques to speech, we think of speech as a doubly stochastic process where the states correspond roughly to acoustic-phonetic events (or to configurations of the human vocal apparatus). We cannot directly observe these states (hence, we call them hidden), but instead, we observe some parameters of the speech waveform such as LPC coefficients, cepstral coefficients, or filter bank outputs. These parameters, in turn, are effectively modeled as random variables whose probability distributions depend on the current state.

To train an HMM recognizer it is necessary to create a model for each word from N training utterances of each word. Each utterance is represented as a sequence of observations. We wish to find that model which has the highest likelihood of producing those sequences of observations. An iterative training procedure, known as forward-backward training,[3,12,13] is used. Initial estimates of the model are made, i.e., the state transition matrix and state observation distributions. Then the probability of the observation sequences are then calculated given the initial model. The parameters are then re-estimated in such a way that the probability score increases. This is done repeatedly until the increase in the probability score is below some threshold value.

During recognition each word model is used to determine the most likely segmentation (or node assignments) for each of the observations, and to calculate the likelihood of the observations for this optimal state assignment given the word model. The word that corresponds to the model producing the highest likelihood score is then chosen as the recognized word. This technique is known as Viterbi decoding.[3,12,13]

There are many complications in the design and implementation of such systems. For example, no theory exists to determine the exact form of the model to use, and it is not clear that the model shown in Figure 2-3 is adequate. In addition, the best number of nodes to be used in a model cannot be determined theoretically. Also, the forward-backward training algorithm can

only guarantee a locally optimal and not a globally optimal set of parameters for the word models it generates.

However, good results have been obtained with simple HMM structures of the form indicated in Figure 2-3, and the system used here as the first-stage recognizer was of this form. An 11-node model with variance limiting was used for all words. Variance limiting constrains the range of variance estimates used in the word models, and is further described in Reference 8.

3. THE LINCOLN STRESSED-SPEECH DATA BASE

3.1 DESCRIPTION

The following is a description of the data base used for all experiments with the discriminant system. It includes speech of subjects under workload stress, as well as speech of subjects instructed to speak in a variety of styles designed to produce the kinds of acoustic variations typical of real physical and psychological stress conditions. The data base was recorded at Lincoln Laboratory in a quiet room and stored on Scotch 3M-208 audio recording tape. A total of nine speakers were recorded. Each speaker went through three, approximately one-hour recording sessions. Approximately one week elapsed between the first and second sessions, and approximately one month between the second and third. The first two sessions comprised the style portion of the data base, and the third session comprised the stress portion.

The vocabulary consisted of 35 aircraft words with many highly confusable subsets. Words were selected from a set of 105 used in a vocabulary created by the Texas Instruments Speech Research Group.[11] The vocabulary was: break, change, degree, destination, east, eight, eighty, enter, fifty, fix, freeze, gain, go, hello, help, histogram, hot, mark, nav, no, oh, on, out, point, six, south, stand, steer, strafe, ten, thirty, three, white, wide, zero.

During the first and second recording sessions, the subject began by saying each word in the vocabulary five times in succession. The subject was then instructed to go through the entire vocabulary eight more times. Each of these times the subject was instructed to speak in a specific manner as listed below.

slow	The speaker was told to say each word slowly.
normal	The speaker was told to say each word normally.
fast	The speaker was told to say each word fast.
soft	The speaker was told to speak softly.
question	To achieve a rising pitch throughout the word, the speaker was instructed to say each word with a questioning intonation.
loud	The speaker was told to say each word loudly.
clear	The speaker was told to say each word clearly and distinctly.
angry	The speaker was instructed to speak angrily.

The third session began in a similar way. The subject repeated each word in the vocabulary twice, in a normal manner, as opposed to five times in the other sessions. The subject was then evaluated on a critical tracking workload task that induces stress.[14] In this task the subject was seated in front of a computer terminal. The subject moves a large triangle on the screen by rotating a knob. The object is to rotate the knob so that the triangle stays on the screen, while it

becomes increasingly difficult to keep the object on the screen. From this initial performance, the maximum difficulty level at which the subject can still maintain control of the task was estimated.

While performing the task at a difficulty level set at 50 percent of maximum, the subject was asked to repeat the words in the vocabulary. This test was performed twice, and then two more times at a difficulty of 70 percent of maximum. These conditions are referred to as co50, and co70. Finally, the speaker was asked to go through the vocabulary while listening to background noise through headphones at a level of 90 dB spl. This condition is referred to as the Lombard condition,[15] and is known to produce substantial acoustic phonetic changes.[5]

All the recorded data were digitized using the SPIRE system,[16] running on a Symbolics 3600 Lisp Machine. All word tokens were low-pass-filtered at 8 kHz before digitization to prevent aliasing. The speech was sampled at 16 kHz using 16-bit samples.

Storage requirements for the data base were very large. For each speaker there were 1190 tokens with an average duration of approximately one-second each. This required 35 Mbytes per speaker, or over 315 Mbytes for the entire 10,710 token data base.

To enable a more efficient environment for running recognition experiments, the data base was also stored in a compressed form. Taking 10 ms as a speech frame, 17 cepstral coefficients were computed for each frame. Two bytes were used to store each coefficient so that an average-length word would require 3400 bytes of storage. The entire data base in this form resides on 34 Mbytes of disk space.

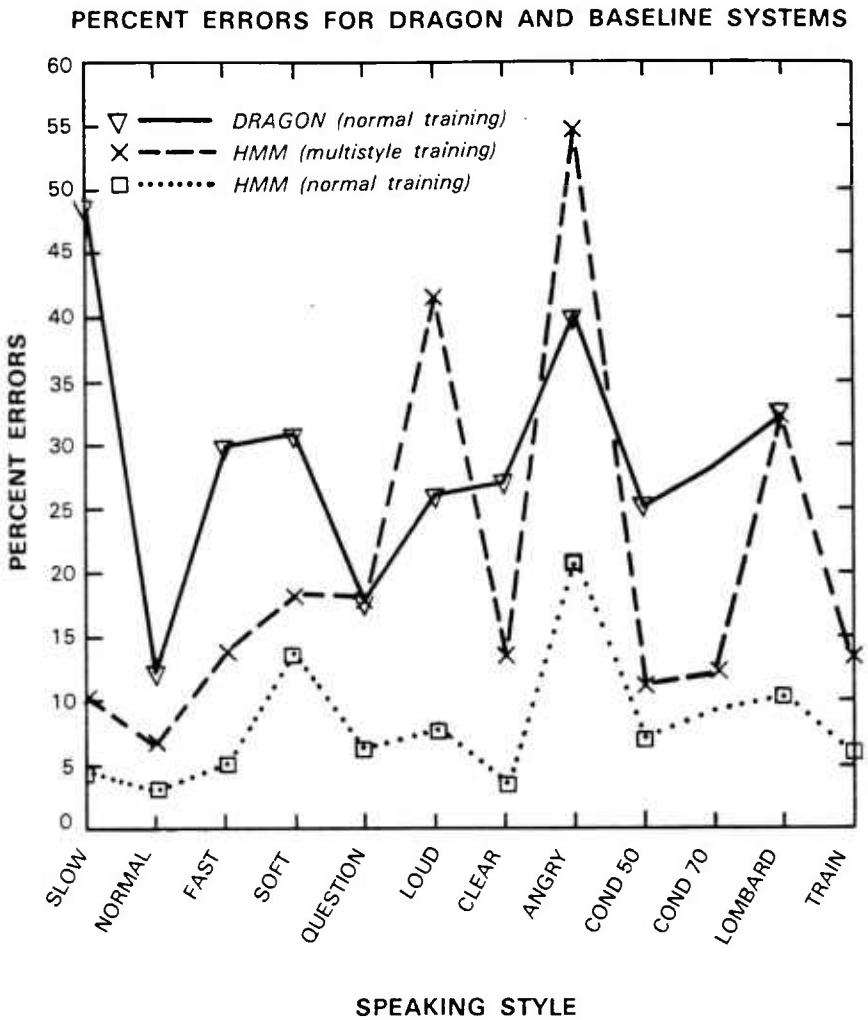
3.2 PRELIMINARY TESTING

3.2.1 The Dragon System

Before any of the Lincoln stressed-speech data base was digitized, a recognition experiment was performed using a Dragon recognition system (software-based) installed on an IBM Personal Computer.[17] The rejection threshold of this system was set to reduce the number of rejections to the point when the system could be effectively characterized in terms of its substitution error rate only. The purpose of this experiment was to provide a comparison of the baseline system to a good commercially available recognizer.

The Dragon system was first trained using the training utterances of the first session (the five normal utterances that begin the session). The system was tested using all utterances from the remaining part of the first session. This amounted to 280 words for each speaker. The Dragon system was retrained in a similar fashion with the training utterances from the second session. The remaining utterances from the second session were then tested. Finally, all the utterances from the third session were tested with the templates generated from the second session.

This testing was unique in that the data base was one consisting primarily of stressed speech. This also gave a valuable initial insight about how a state-of-the-art commercial recognizer, trained with normal speech, would behave with such a data base. The solid line of Figure 3-1



78054-10

Figure 3-1. Comparison of the Dragon and baseline systems. The solid line shows error rate for the Dragon system, the dashed line shows error rate for the baseline system using normal training, and the dotted line shows error rate for the baseline system using multistyle training.

describes the resulting error rates of this experiment. The horizontal axis spans the eight speaking styles, the easier (co50) and more difficult (co70) workload stress conditions, the Lombard condition (Lomb), and the remaining tokens not used during training. It includes all the styles and conditions found in the data base.

The error rate of the Dragon recognizer, when trained on normal speech and tested on the stressed-speech data base, is much higher than could realistically be tolerated in any real application. This reveals a serious problem with recognition of stressed speech. The overall average error rate on the entire data base was 28.5 percent. The error rates were lowest for the normal and

question conditions, and highest for the slow and angry conditions. It may be speculated that the question and normal conditions are the conditions acoustically closest to the training condition, and the slow and angry conditions are the conditions most acoustically different from the training.

After examining the frequency of specific confusions, nine subvocabularies were made. The size of these vocabularies ranged from two to six words. The subvocabularies were as follows:

subvocabulary 1	degree, three, thirty, fifty, freeze
subvocabulary 2	eight, eighty, gain, change
subvocabulary 3	east, fix, six
subvocabulary 4	go, hello, no, oh
subvocabulary 5	enter, ten
subvocabulary 6	white, wide, point, break, strafe
subvocabulary 7	help, out, south, hot, zero, on
subvocabulary 8	steer, stand, destin
subvocabulary 9	mark, nav, histog

Most confusions were within, but not between, these groups. While the groupings only account for 5.48 percent of all possible pairwise confusions, the percent of all errors that fell into these groups was 51.35 percent.

Two conclusions are drawn from this experiment. The first conclusion is that, trained in this manner, the performance of the Dragon system was inadequate. More sophisticated training and recognition algorithms must be used to compensate for stressed speech. The second conclusion is drawn from the observation of the errors falling into subvocabularies quite frequently. This is an advantageous condition for the performance of two-stage discrimination. Given this observation, the use of a two-stage system appeared viable.

3.2.2 The HMM Baseline System

After completely digitizing the data base, a similar experiment was conducted using the HMM baseline recognition system. One training set of five word tokens was used for each speaker to create one word model to test all other tokens. Two training techniques were used. First, five tokens of normal speech were used. Then, multistyle training[8] was used where the five tokens, all taken from the first session, consisted of one token spoken as part of the initial training set, one token spoken fast, one token spoken with a questioning intonation, one token spoken loudly, and one token spoken clearly. The system used was the top performing HMM system at that time on the Lincoln data base. A number of experiments conducted at Lincoln[5,8] have shown multistyle training to be very effective in improving recognizer performance.

Results from testing the HMM recognizer with multistyle training (one token of normal, fast, question, loud, and clear speech) and normal training (five normal tokens) are presented in Figure 3-1, along with results from the Dragon system. The dashed line with crosses shows the error rates for the baseline system using normal training. The overall error rate of this system is 18.9 percent. It performs better than the Dragon system for all conditions, excluding angry and loud. Results of multistyle training are presented by the dashed line with squares of Figure 3-1. The overall error rate fell from 18.9 percent with normal training to 7.7 percent with multistyle training. This illustrates the advantage of multistyle training, and the basis for selecting it for use in these experiments.

It should be noted that recognition performance of the Dragon system with multistyle training was not tested. It would be expected that multistyle training would also improve performance for the Dragon system. The objective here, however, was not to evaluate any particular commercial system but to establish a high-performance baseline HMM system as a basis for the discriminant analysis work. The results in Figure 3-1 are intended to establish that the Lincoln HMM system with multistyle training provides such a baseline.

As mentioned above, a scheme was needed to determine under which circumstances a second-stage analysis would be used. If a two-stage procedure is always used, a risk would be run in degrading the performance of an already tested and effective first-stage system. By never using the second stage, the problems mentioned earlier with the baseline system would remain. A rule was therefore sought that would only use the second stage when the probability of the first stage producing an error was high. A discussion of the rules used in the experiments performed is found in the next section.

A statistic that proves to be crucial in determining an upper limit on the performance of a two-stage system is the percent of all trials that result in the correct word being in the top few candidates. If the correct word is not among the top few candidates in a large percentage of cases where the baseline system was incorrect, then the discrimination system can offer no improvement. If the correct word is always among the top few candidates, the discrimination system has the potential to correct all errors. The upper limit on the performance of a two-stage system is the percent of all trials that includes the correct choice in the top N selections, where N is the number of candidates the second stage uses. Table 3-1 demonstrates that by discriminating among just the top two candidates, the error rate could potentially be halved. By performing a three-way discrimination the overall error rate could be reduced to as little as 1.6 percent.

TABLE 3-1

Percent Errors in Top N Selections

Speaker	1	2	3
g1	18.7	8.4	4.9
g2	5.5	2.0	1.3
g3	7.1	2.8	1.8
n1	4.9	1.1	0.5
n2	9.4	4.7	2.9
n3	4.8	0.6	0.1
b1	8.6	3.0	1.6
b2	6.1	2.2	0.6
b3	3.3	1.4	0.8
Total	7.7	2.9	1.6

Percent errors for each speaker out of 1015 tokens. Error rates are for the top one, two, and three candidates from the baseline system.

4. DISCRIMINANT ANALYSIS

Baseline training provides word models for each word in the vocabulary based on five training utterances. These models are obtained with the use of forward-backward training described earlier. This training technique repeatedly passes the training utterances for a given word through that word's model using the forward-backward algorithm. Training required for the discriminant system involves passing all training utterances through all word models using the Viterbi decoding scheme.

Suppose word models A and B produced essentially equal likelihood scores from the baseline system for a given input sequence. In other words, the input token's behavior, when passed through model A, matched closely to the way A's training set behaved when passed through the same model. This match was essentially equal to the match between the way the input token behaved when passed through model B, and the way B's training set behaved when passed through model B.

With the existing baseline system, a decision must be made at this point, based on limited information obtained by passing training tokens through their own word models. With the discriminant system more information is available. With the discriminant training data the key differences in the words A and B are brought out. This is done using the statistics obtained from passing training instances of both of these words through both word models. This is illustrated by the following example.

Define an observation sequence for an unknown word as O.

$$\underline{O} = \{\underline{O}(1), \underline{O}(2), \underline{O}(3),...,\underline{O}(n)\} \tag{4.1}$$

There are n 10-ms time frames in the unknown observation sequence. The t^{th} observation frame consists of 16 cepstral coefficients.

$$\underline{O}(t) = \{O_1(t), O_2(t),...,O_{16}(t)\} \tag{4.2}$$

When decoded on a word model using Viterbi decoding, the observation sequence is segmented by observing the backtrace which assigns each observation frame to a state. The optimal state transitions that were determined during decoding and used for segmentation are then recorded. (Many applications of this type of decoding make no use of this backtrace, since just the optimal score that results is of interest.) Each segment is assigned to a node. For instance, let us assume there are four nodes in a particular word model.

node 1 $\{\underline{O}(b_1)....\underline{O}(e_1)\}$

node 2 $\{\underline{O}(b_2)....\underline{O}(e_2)\}$

node 3 $\{\underline{O}(b_3)....\underline{O}(e_3)\}$

node 4 $\{\underline{O}(b_4)....\underline{O}(e_4)\}$

Observations b_1 through e_1 are assigned to node 1. Observations b_2 through e_2 are assigned to node 2, etc. Frames e_i and b_{i+1} will always correspond to adjacent observations. As part of the

discriminant training. all the training utterances are decoded this way. Statistics are formed for all training tokens of every word using each word model, word, node, and parameter. Parameters used included 16 cepstral coefficients and relative energy, and two node duration parameters. Also used were 17 differential feature parameters, corresponding to the node-to-node difference in mean value for the 16 cepstral parameters and energy. The following describes how these statistics are generated.

Statistics for word B being tested with word model A are recorded. These statistics are generated from the Viterbi segmentations of the word B training utterances when tested by model A. Each segmentation is a set of N mean model-parameter vectors, where N is the number of nodes in the model.

The means and variances of each model-parameter vector are estimated as an independent Gaussian random variable for each word, word model, and node. Given a training set of observations, statistics are obtained for that set with any word model. Let $\underline{S}_i$ be the mean observation vector for node i.

$$\underline{S}_i = \frac{1}{e_i - b_i + 1} \sum_{k=b_i}^{e_i} \underline{O}(k) \tag{4.3}$$

$\underline{O}(b_i)$ is the first observation vector in the segment, and $\underline{O}(e_i)$ is the last observation in the segment. For this given input token, a mean model-parameter vector for each node in the model is obtained.

Therefore, if we are given five utterances to use for training, a mean observation vector can be obtained for the input word, the model, and the node number of the model. For each training utterance and word model, a set of S_i's are obtained. Let $\underline{S}_{iq}$ be the mean observation vector for the q^{th} training utterance in node i. The mean observation vector, given these utterances, is just the simple mean of these utterance mean vectors for each node.

$$\underline{S}_i \text{ for five training utteranes } = \frac{1}{5} \sum_{q=1}^{5} \underline{S}_{iq} \tag{4.4}$$

Similarly, from these five samples the variances are estimated. These statistics are stored as four-dimensional arrays whose indices are word model, word, node, and model-parameter.

Assume an observation sequence $\underline{O}$. After training and testing on a baseline system it is determined that no decision can be made from the baseline alone, although the decision should clearly be either word A or word B. The discriminant system must be used to decide between these two candidates.

Word model A is first studied. A statistical description of an utterance of word A being passed through model A is used. Also, a similar description of an utterance of word B being passed through model A is used. This statistical description comes from passing the training tokens of word A through model A, and the training tokens of word B through model A. For

instance, Figure 4-1 below examines node i of model A when the utterance is A and when the utterance is B. For simplicity, it is assumed that only three model parameters comprise the vector.

Figure 4-1 illustrates the contribution to the final likelihood ratio of a single node. For the above example the distribution for each of the model-parameters is given. There exists a set of distributions for the case of the input word being A and a set for the input word being B. The assumption made is that the segmentation has arisen from model A. By plotting the observation vectors for the input word in node i against these distributions, the likelihood ratios between

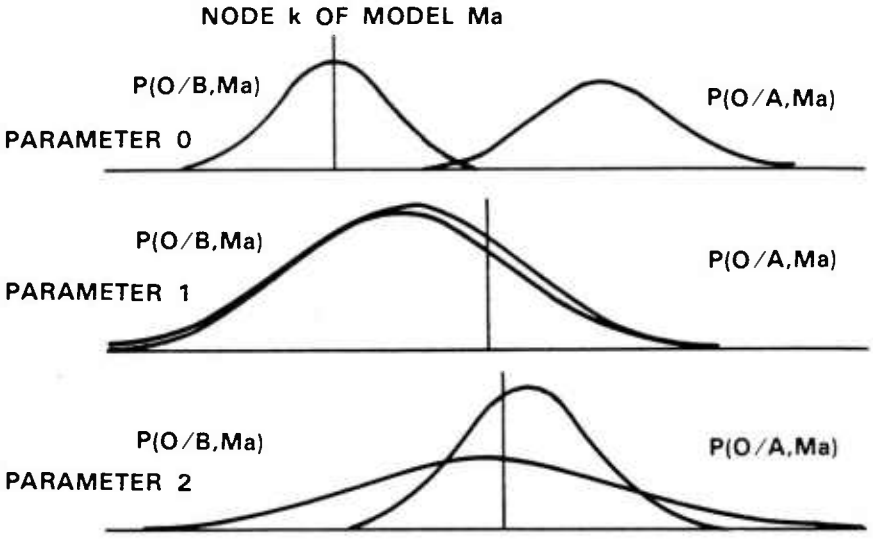


Figure 4-1. Illustration of likelihood-ratio scoring; parameter 0 appears to be the only clear discriminator between A and B.

words for each model-parameter can be found for that node. It should be noted that in Figure 4-1 only parameter 0 can be thought as a good discriminator. By looking at the likelihood ratios from all of these parameters it is clear that parameter 0 will have the strongest effect. The classifier focuses attention toward that parameter which is the best discriminator. This is a crucial feature for a discriminator to perform well on words that are acoustically similar.

Given that we have two candidate choices (A and B), an unknown observation sequence $\underline{O}$, and the two respective word models used for segmentation M_a and M_b, we wish to calculate two statistics.

$$L_a = \log \frac{P(\underline{O}/A,M_a)}{P(\underline{O}/B,M_a)} \tag{4.5}$$

$$L_b = \log \frac{P(\underline{O}/B,M_b)}{P(\underline{O}/A,M_b)} \tag{4.6}$$

If L_a and L_b are of opposite sign, then the two scores both favor the same word. Each of these statistics is the sum of the individual contributions from each node. Let $\underline{O}$ (in node i) be the set of observation vectors $\underline{O}$ (b_i) through $\underline{O}$ (e_i), or all the observations assigned to node i.

$$\log P(\underline{O}/A,M_a) = \sum_{i=1}^{N} \log P[\underline{O} \text{ (in node i)}/A,M_a] \tag{4.7}$$

Similarly, each node score is the accumulated result of the individual contributions from each observation assigned to that node.

$$\log P[O \text{ (in node i)}/A,M_a] = \frac{1}{e_i - b_i + 1} \sum_{j=b_i}^{e_i} \{\log P[O(j)/A,M_a]\} \tag{4.8}$$

It is important that this score is normalized by the number of observations assigned to the node. Failing to do this will introduce a durational dependence into the scoring. This will have the effect of giving unequal weight to nodes with different numbers of observations assigned to them. However, a legitimate reason for calculating scores this way (per-observation) would exist if the observations were uncorrelated.

The claim here is that they are correlated and the units that should be treated as uncorrelated events are the nodes rather than the observations.

4.1 USE OF ADDITIONAL FEATURES IN DISCRIMINANT ANALYSIS

To enhance the discriminant analysis, it may be desirable to include additional features, either on a per-observation frame, or on a per-node basis. Discriminant training, as described above, is used to estimate means and variances of these new features for each input word, word model, and node. To incorporate a new per-observation feature we use the equation

$$\log P[\underline{O} \text{ (in node i)}/A,M_a] = \frac{1}{e_i - b_i + 1} \sum_{j=b_i}^{e_i} \{\log P[\underline{O}(j)/A,M_a] + \log P[f(j)/A,M_a]\} \tag{4.9}$$

where f(j) is the value of the feature at observation frame j. To include a per-node feature, that is a measurement such as node residency time which is computed once for each node, we use:

$$\log P(\underline{O}/A,M_a) = \sum_{i=1}^{N} \{\log P[\underline{O} \text{ (in node i)}/a,M_a] + \log P[f(i)/A,M_a]\} \tag{4.10}$$

where here f(i) represents the value of the measured feature in node i.

4.2 THE DURATION MODEL

The node residency time statistic has been used as a per-node feature to model the duration of each HMM state. It represents the only explicit inclusion of nodal duration in the discriminant analysis, since per-observation features are normalized by averaging over each node.

Two duration models were implemented. The first, known as absolute-duration, is based on the actual number of observations assigned to a node. The second, known as relative-duration, is based on the fraction of the entire word spent in any node. This relative-duration model, used previously in Reference 15, would be most effective under the assumption that, as the overall length of a word is increased or decreased, the individual acoustic events are increased or decreased proportionately. In either case, the values seen during recognition, for the duration statistic, are simply included as an added feature, as described in the previous section with Equation (4.10).

4.3 ESTIMATION PROBLEMS DUE TO LIMITED TRAINING DATA

A large number of statistics are estimated when training the discriminant system described above. Many of these statistics are estimated from sample sets as small as five. The estimation error, resulting from these small sample sets, must be considered to avoid degraded performance.

Variance estimates are much more sensitive to the effects of limited training data than are mean estimates. A possible solution is to eliminate the use of variance from the likelihood-ratio calculation. This could be used in situations where only a single model-parameter is being used in the discriminant system. In situations where multiple model-parameters are used, variance must at least be included to effectively weight the likelihood ratios obtained from different model-parameters. By calculating a variance taken from samples across all words, nodes, and models, a statistic derived from a much larger sample set is achieved. This scheme assumes that the variances of the same model-parameter, given different hypotheses (word A or B), are identical. Equation (4.11) shows the calculation involved in generating the likelihood ratio from a single observation vector in node k, using the grand variance of each of the parameters:

$$\log \frac{P[\underline{O}(t) \mid A, M_a]}{P[\underline{O}(t) \mid B, M_a]} = \sum_{i=o}^{\text{number of parameters}} \frac{[O_i(t) - \mu_{A,a,k,i}]^2 - [O_i(t) - \mu_{B,a,k,i}]^2}{var_i} \tag{4.11}$$

In this equation $\mu_{A,a,k,i}$ is the estimated mean value of parameter i in node k, given that the input is word A being segmented by model A. $\mu_{B,a,k,i}$ is the estimated mean value of parameter i in the node k, given that the input is word B, being segmented by model A. By assuming the distributions to be Gaussian, this computation is made efficient.

The peformance of recognition systems often degrades when many features used for discrimination are added. Fundamentally, the adding of more information, such as the information a feature gives, should not degrade performance. The reason the contrary has been observed is primarily a result of small sample sets used to estimate statistics. A nasality detector can be a powerful feature. When discriminating between "pop" and "top" a nasality detector can be useless. If infinite training data were available, a nasal feature of this sort would probably not contribute a great deal to a likelihood-ratio test. This is simply a result of the nasal discriminator not finding the feature it was meant to extract. Again, this presents no problem if the training data were infinite. The contribution of this discrimination in a likelihood-ratio sense would just be very small. Small sample sets from training do present a problem in this case. Since estimates such as variance in the nasal detector will be very poor in these situations, the decision boundaries produced by the discriminator can be altered from their true value enough to degrade the overall effectiveness of the system. Using such a feature in all discriminations could add enough statistical noise to cases where the feature is irrelevant to completely overwhelm any useful information obtained from other model-parameters in different nodes.

In general, when only limited training data are available the addition of many features into a recognizer is usually accompanied by the addition of a great deal of statistical noise in the scoring. A solution to this problem is to use only those features that are appropriate in a discrimination, given the words and nodes that are to be discriminated.

Statistical T-testing is a method that addresses this problem.[18-20] Estimates made from small sample sets can be poor. Two distinct sample sets may exhibit distinct estimates even though the underlying statistics for these two sets are identical. Given this situation, the two sets of estimates should not correspond to separate categories in a classifier. A classifier should have some mechanism for determining whether estimates correspond to identical underlying statistics. The statistical T-test provides a mechanism for doing this.

$$T = \frac{\hat{x} - \hat{y}}{(n_x - 1)\,S_x^2 + (n_y - 1)\,S_y^2} \sqrt{\frac{n_x n_y (n_x + n_y - 2)}{n_x + n_y}} \tag{4.12}$$

Equation (4.12) defines the parameter used in the T-test. $\hat{x}$ and $\hat{y}$ are the estimated means of a feature (in a given node, for a given word model) for words A and B. S_x^2 and S_y^2 are the estimated variances. n_x and n_y are the number of samples of x and y, respectively. This formula assumes that the variances of the underlying features are equal. This assumption was confirmed using F-ratio tests with the estimated variances during preliminary experiments. If this assumption had not been confirmed, a similar test from Satterthwaite[21] which allows unequal variance could have been used.

To use this test a value of T is compared to a threshold, so that if the absolute value of T is greater than the threshold value, the two statistics are accepted as having different means. If the absolute value of T is less than this, the two statistics are accepted as having identical means. If

the latter is the case, that particular model-parameter will be discarded from the scoring procedure for that node. It is suspected that a confusion-specific feature will have a small mean difference relative to the variance in situations where the feature is not relevant. This will allow the statistical T-test to remove it from the scoring scheme, thus eliminating a major cause of statistical noise and degraded performance. The threshold for all T-tests can be adjusted for various significance levels. For experiments, this threshold was adjusted so that the probability of a false acceptance is 0.05, where a false acceptance is defined as classifying estimates with identical underlying statistics as having different underlying statistics.

4.4 DECISION SCHEMES

Techniques described thus far in the discriminant system yield two final statistics from which a decision must be based. Those statistics are defined in Equations (4.5) and (4.6). Three schemes were tried in experiments for deciding on a selection. The first scheme was to simply take the difference of these two statistics and choose B if the difference is >0, and A if the difference is <0. This scheme is depicted in Table 4-1. Another scheme was to only accept the second stage's result if the two statistics agree with each other. This is equivalent to the two statistics being of opposite sign. This only uses the second stage in cases where it has made a strong decision. In cases where the statistics do not agree the decision was deferred back to the original baseline score. This is depicted in Table 4-2. A problem with the above scheme is the possibility for a very large number of deferrals to occur.

A third scheme is presented which tries to decrease the frequency of deferring. A threshold is decided based on the absolute value of the difference between the statistics. If this threshold is exceeded, at least one of the statistics must point strongly to a solution. Then, a decision is made based on their difference. If this threshold is not exceeded, then the decision is again deferred to the baseline system. This scheme has the advantage of not forcing a decision if both statistics are close to zero. This scheme is depicted in Table 4-3.

TABLE 4-1

Nondeferring Pairwise Decision

$L_a - L_b$	Decision
>0	A
<0	B

Decision scheme does not allow deferring. Candidate with best overall score will be selected as the recognized word.

TABLE 4-2

Deferring Pairwise Decision

L_a	L_b	Decision
>0	>0	defer
>0	<0	A
<0	>0	B
<0	<0	defer

Decision scheme allows for deferring. It only accepts the second-stage decision if L_a and L_b both point to the same word.

TABLE 4-3

Limited Deferring Pairwise Decision

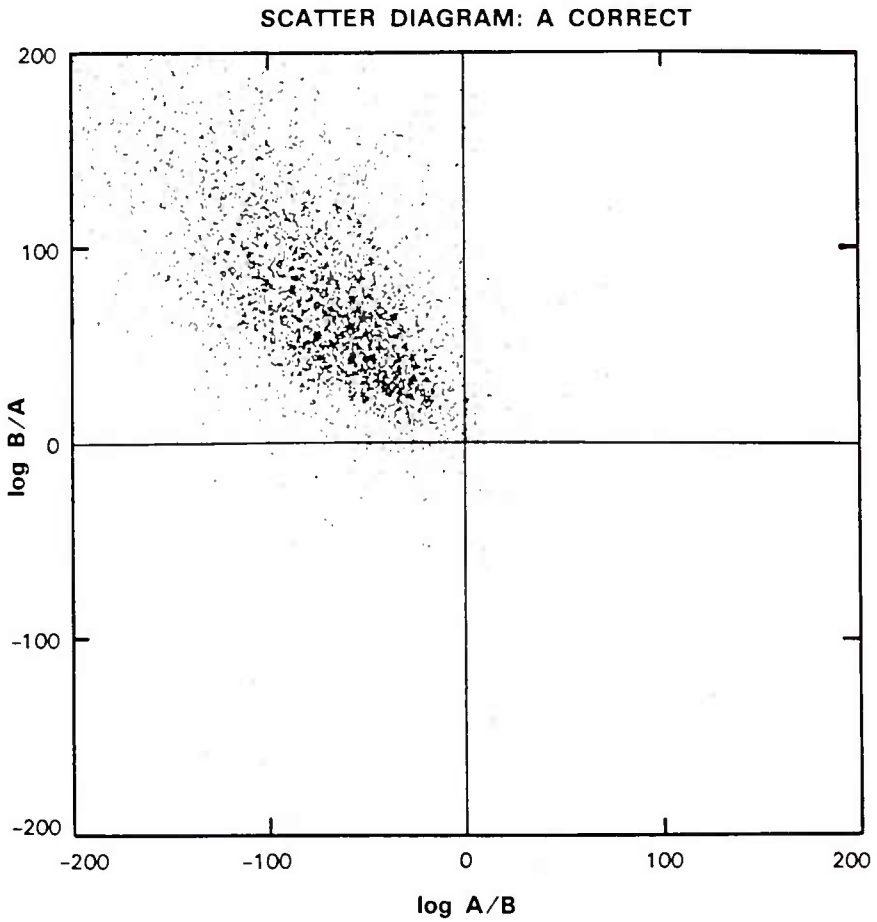
$L_a - L_b$	Decision
>T	A
<-T	B
else	defer

Decision scheme to limit number of deferrals. This will accept the second stage in some cases where the two scores point to different words. If one points strongly to a word, and the other points weakly to the other, the word could be accepted depending on threshold T.

Figures 4-2 and -3 show the scatter of points corresponding to these two scores; L_a and L_b measured for all test words in the data base. Figure 4-2 is a scatter diagram for all tokens where the baseline chose the correct word. Figure 4-3 is a scatter for all tokens that the baseline's second choice was correct. Both plots are scatters of the scores obtained from the second stage. The clustering into the upper left and lower right quadrants in these figures suggests that these scores will be useful in discriminating.

4.5 WHEN DISCRIMINATION IS NEEDED

An assumption made throughout this report is that the first-pass system is a good system with a reasonably low error rate. Given this assumption, the second-stage system should not be



78054-2

Figure 4-2. Scatter of scores from the second stage. Horizontal axis represents the likelihood ratio from model A. Vertical axis represents the likelihood ratio from model B. Points to the top and left represent scores strongly favoring word A. Points to the right and bottom represent scores strongly favoring B. All points correspond to instances where A was the correct word and selected as the first choice by the baseline. Ninety-five percent of all points are found in the top left quadrant.

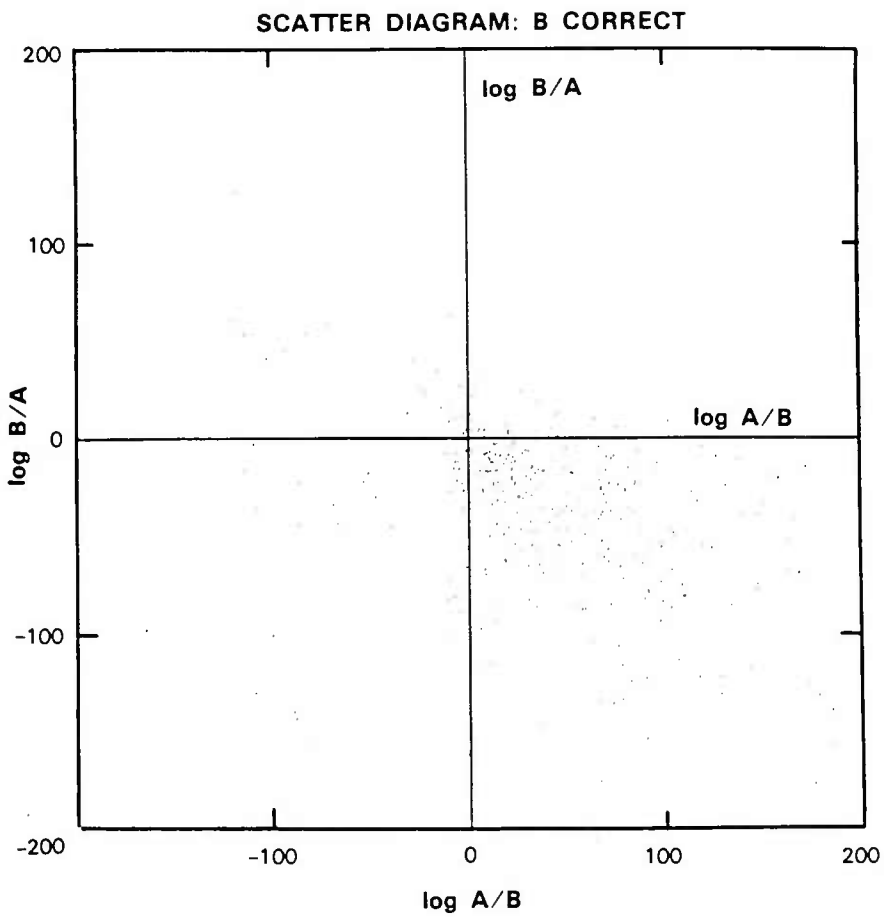
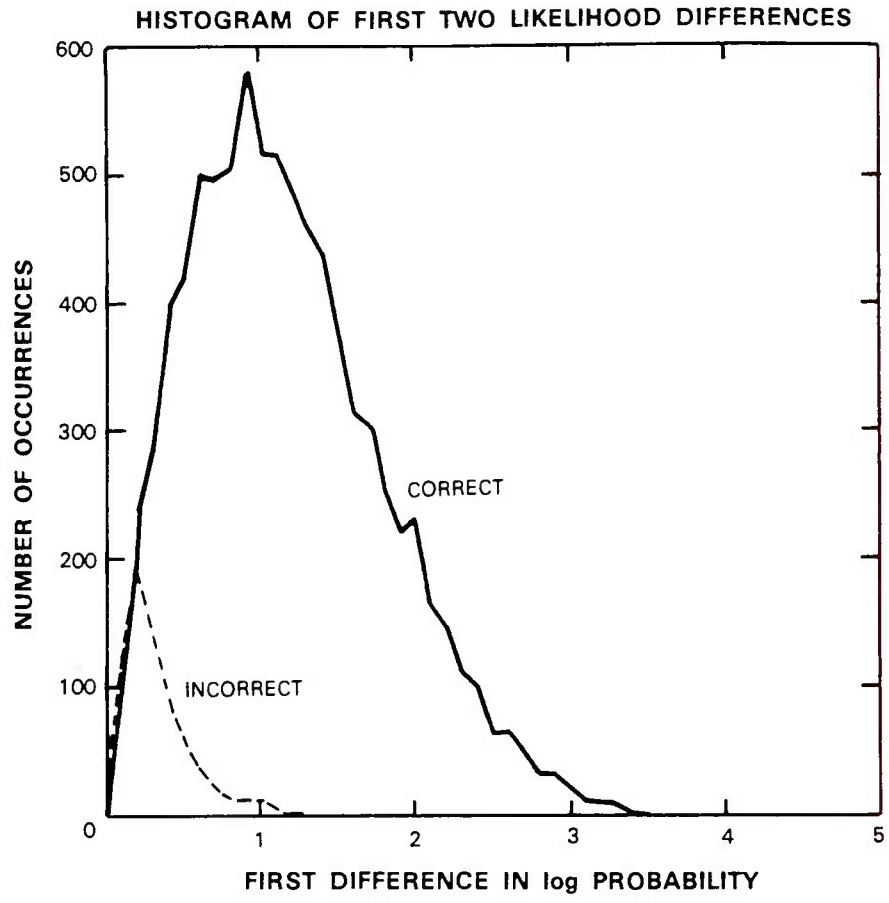


Figure 4-3. Scatter from second stage, similar to Figure 4-2. All points correspond to instances where B was the correct word chosen as the second best candidate from the baseline. Seventy-seven percent of all points are found in the bottom right quadrant.

needed for all input words. It should only be used when the probability of the first pass producing an error is great. A threshold is set that serves to identify two regions. They are: the region where first-stage errors are likely, and the region where these errors are not likely. This threshold is based on the difference between the top two scores from the first-pass system.

Figure 4-4 shows how these regions can be identified. Virtually all first-stage errors result in the difference between the top two baseline scores being <1.5. The thresholds that were used in all experiments were 0.5 and 1.0. With the threshold set at 0.5 all but 79 words recognized incorrectly by the first stage would be discriminated. With the threshold set at 1.0 all but 9 of these would be sent to the second stage.



78054-8

Figure 4-4. Histogram of first difference of scores from baseline system. Horizontal axis plots the difference in log-probability scores for the top two candidates. The vertical axis plots the number of occurrences. The two lines correspond to histograms of correct responses and incorrect responses by the baseline.

4.6 DISCRIMINANT DECISION FLOW

The following serves as a description of exactly how an unknown word passed to the two-stage system is processed, and either successfully or unsuccessfully recognized.

An unknown word is passed through the first-stage recognizer. If the correct word is not among the top N candidates, the system will have no chance of obtaining the proper solution. This case will result in an error. If the correct word is among the top N candidates, the correct result may be obtained. If the difference between the first two scores is greater than the set threshold, the baseline score will be used to select a word. If the difference is less than the threshold, unknown input is passed to the discriminant system. The discriminant system either selects one of the top two candidates as the recognized word or defers back to the baseline score. This scheme is depicted in Figure 4-5.

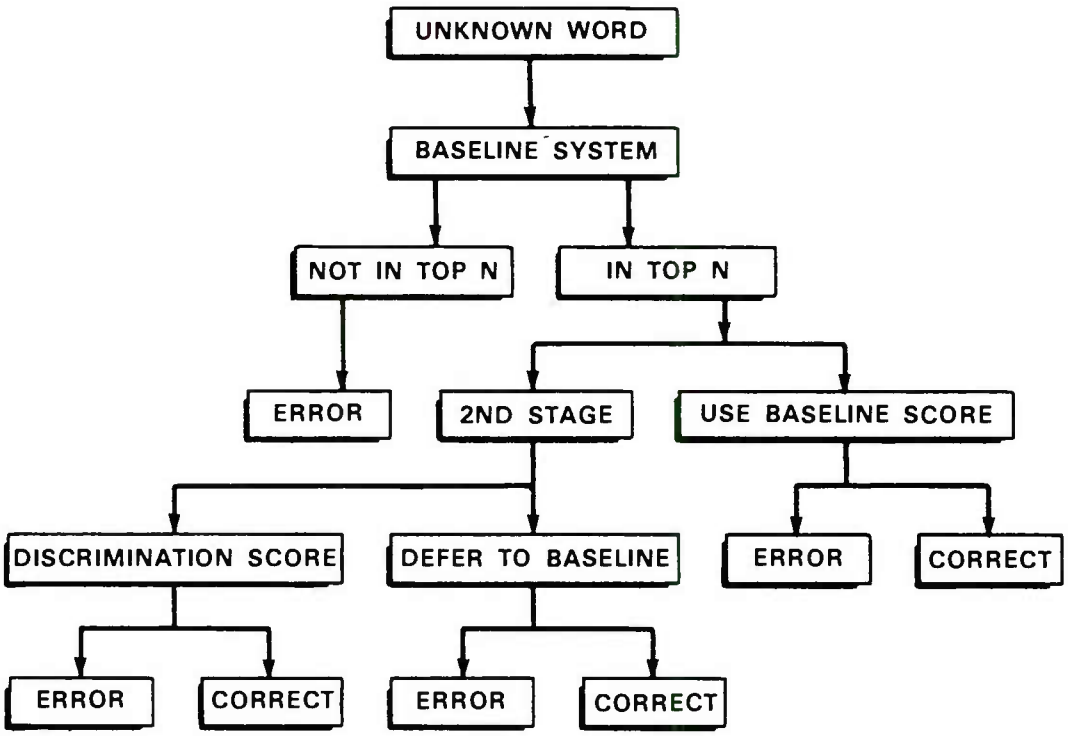


Figure 4-5. Flow diagram of decision logic used in the two-stage recognizer.

5. EXPERIMENTS AND RESULTS

The following is essentially a chronological description of the most important experiments performed with various forms of the discriminant system. Outlined are descriptions of experiments which reduce the 7.7-percent error rate of the baseline system to 3.5 percent. Tables 5-1 through 5-3 give a detailed look at the results of these experiments. Table 5-4 gives a more detailed look at four of the best systems. It describes the decision flow of the input words and reveals which circumstances result in the remaining errors. Figure 5-1 gives a graphic comparison of the HMM baseline system and the best two-stage system from experiment 10. As can be seen from this figure, discriminant analysis is effective for all stress conditions and talking styles.

Experiment 0 — The Baseline System

This experiment is designed primarily to evaluate the use and performance of the first-stage recognizer, as well as to collect segmentations from the Viterbi decoding. The recognizer was an

TABLE 5-1

Features Defining Each Experiment

Modification	Experiment												
	0	1	2c0	2dur	3c0	3dur	4both	4eith	5	6	7	8	9	10
c1-c16	X	X							X	X	X	X	X	X
c0		X	X		X		X	X	X	X	X	X	X	X
Abs_dur		X		X		X	X	X	X	X	X	X	X	X
T-test										X	X	X	X	X
Features											X	X	X	X
Fixed Var					X	X	X	X	X	X	X	X	X	X
Threshold 1.0												X	X	X
Threshold 0.5	X	X	X	X	X	X	X	X	X	X	X			
Top 3													X	X

The "X's" correpond to a particular modification being included with a particular experiment. The top row indicates which experiment is being defined. The left column indicates which modifications are used.

TABLE 5-2

Error Breakdown for Each Speaker

Experiment	Speaker										
No. of Tokens	g1 1015	g2 1015	g3 1015	n1 1015	n2 1015	n3 1015	b1 1015	b2 1015	b3 1015	Total 9135	Percent
Experiment 0	193	56	74	51	97	49	89	64	34	707	7.74
Experiment 1	168	58	68	47	110	61	86	49	34	681	7.44
Experiment 2											
c0	199	73	92	85	90	39	87	66	37	768	8.41
dur	167	56	69	42	112	32	80	53	34	645	7.06
Experiment 3											
c0	174	53	64	51	84	35	86	46	28	621	6.80
dur	137	66	61	31	95	37	73	49	33	582	6.37
Experiment 4											
both	157	48	59	35	86	36	75	49	26	571	6.25
either	146	50	58	25	84	31	69	45	31	539	5.90
Experiment 5	116	37	42	29	81	13	50	28	24	420	4.60
Experiment 6	116	36	42	26	82	14	49	25	23	413	4.52
Experiment 7	118	34	44	25	74	16	45	27	23	406	4.40
Experiment 8	111	35	40	27	72	16	39	28	22	390	4.27
Experiment 9	93	35	42	25	70	12	33	18	23	351	3.84
Experiment 10	74	37	34	19	75	10	35	16	25	325	3.56

All errors are broken down by experiment and speaker. Totals are given for each experiment as well as overall percent error rate.

TABLE 5-3

Error Breakdown for Each Style/Condition

Experiment	Speaking Style												
No of Tokens	slow 630	norm 630	fast 315	soft 630	ques 315	loud 315	clea 315	angr 630	50 630	70 630	lomb 630	trai 3465	Percent
Experiment 0	28	18	16	87	20	25	11	134	45	58	65	200	7.74
Experiment 1	42	19	18	87	14	23	14	126	55	48	65	169	7.44
Experiment 2 c0	34	28	27	91	19	29	9	144	54	61	76	196	8.41
dur	42	10	17	82	12	19	11	119	48	51	64	170	7.06
Experiment 3 c0	28	18	20	72	14	28	10	127	45	44	64	151	6.80
dur	33	16	17	74	11	15	10	113	37	43	63	150	6.37
Experiment 4 both	26	16	12	70	12	17	7	117	41	47	59	147	6.25
either	30	13	19	64	11	15	9	112	39	33	58	136	5.90
Experiment 5	22	8	6	56	11	13	4	101	25	26	50	98	4.60
Experiment 6	20	8	8	55	11	14	4	99	25	25	48	96	4.52
Experiment 7	18	7	6	55	9	15	3	98	23	26	47	99	4.40
Experiment 8	18	8	6	53	9	15	3	93	21	26	46	92	4.27
Experiment 9	17	7	6	48	7	12	2	83	21	26	40	82	3.84
Experiment 10	21	8	8	46	10	12	4	84	18	20	31	63	3.56

Errors are broken down by experiment and style/condition. Under each style heading is a number which corresponds to the total number of test tokens.

TABLE 5-4

Decision Flow Statistics for Selected Experiments

	Experiment			
	7	8	9	10
Words Total number of test tokens	9135	9135	9135	9135
Correctable Number of trials in which the correct word was among the top N candidates of the baseline	8873	8873	8991	8991
Sent to 2nd Stage Number of words which were passed to the second stage discriminator	2044	4689	4799	4799
Deferred Number of words sent to second stage which deferred back to the baseline score	249	393	603	34
Errors from Uncorr. Number of errors resulting from the correct choice not being among the top N	262	262	144	144
Errors from Deferring Number of errors resulting from words being deferred back to the baseline	76	82	152	10
Errors from Not Discrim Number of errors resulting from the word not even being passed to second stage	38	5	8	8
Errors from 2nd Stage Number of errors caused directly from the 2nd stage	30	41	47	163
Total Errors Total number of errors in experiment	406	390	351	325

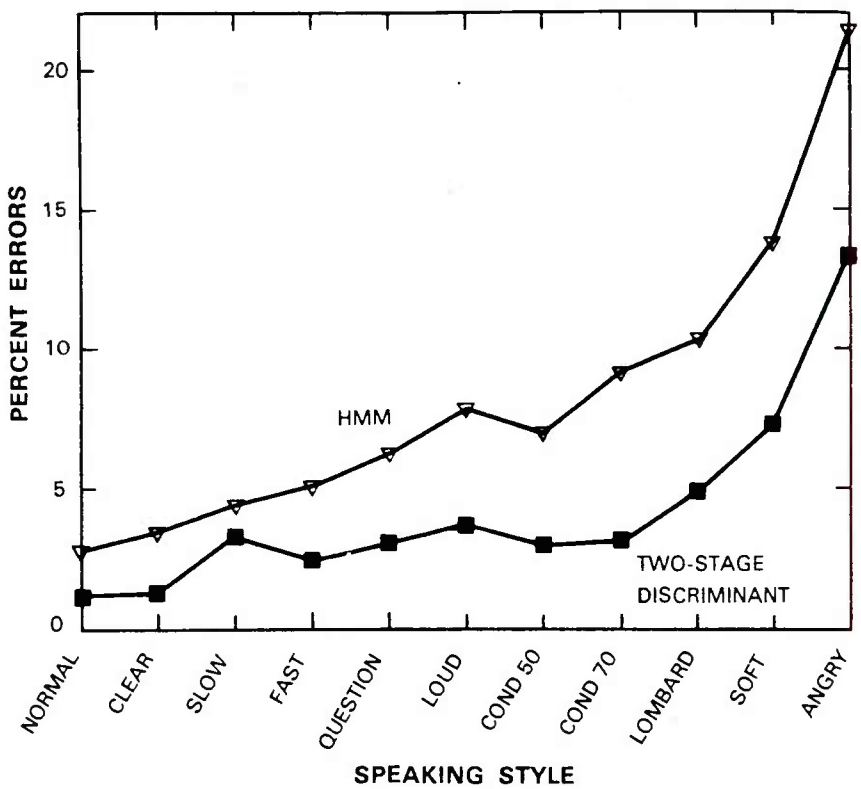


Figure 5-1. Comparison of baseline and best two-stage system. The two-stage discriminant system outperforms the baseline system for all conditions.

HMM system using 11 nodes with variance limiting imposed on the observation distributions. Multistyle training was used. The error rate was 7.7 percent. This corresponds to 707 errors out of 9135 test words.

Experiment 1 — Multiple Model-Parameter Discrimination with Estimated Variances

This was the first attempt at the two-stage discriminant system. The model-parameters used for discrimination were cepstral coefficients c0 through c16 and the absolute node residency time feature. Backtraces from experiment 0 were used for obtaining segmentations. The decision scheme allowed deferring, and was consistent with the method described in Table 4-2. The threshold for using the second stage was set at 0.5. In this experiment and most subsequent experiments, only the top two candidates were used for discrimination. The final error rate of this system did show some slight improvement. The total number of errors was reduced by 27. This corresponds to 680 errors or an error rate of 7.4 percent. At this point the system showed some promise, but significantly better performance seemed to be required to justify the use of the second stage.

Experiment 2 — Isolated Model-Parameter Discrimination

The mediocre results of experiment 1 suggested that perhaps only a few of the model-parameters used for discrimination were in fact effective discriminators. This experiment was then performed to investigate the effects of using each of the model-parameters separately as discriminators. The idea of not being able to defer back to the baseline was also felt worthy of investigation. The experiment involved using only one model-parameter at a time in the second-stage system for discrimination. Each model-parameter was used separately on the entire data base. In addition to the model-parameters used in experiment 1, relative-residency time duration was also used. This experiment proved to be more encouraging. Results showing improved performance over the system in experiment 1 were found. These results are presented in Figure 5-2. The hollow squares indicate points where deferring was not used, and the upside-down triangles indicate

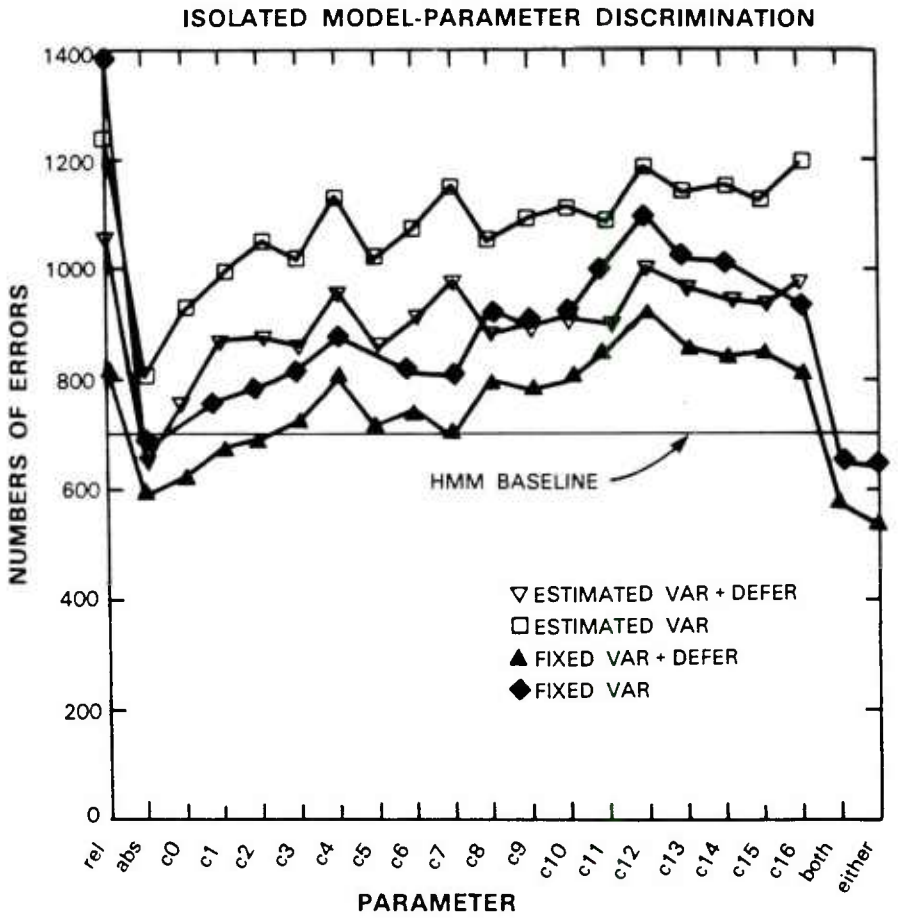


Figure 5-2. Detailed results of experiments 2, 3, and 4. Four sets of data are shown. They refer to both the use of estimated variance and fixed variance, and discrimination with and without deferrals.

points where deferring was used. The horizontal axis indicates the parameters used for discrimination. In this experiment the relative duration (rel), and absolute duration (abs) models were both used, as well as cepstral coefficients c0 through c16. The solid horizontal line shows the error rate achieved by the baseline system. In all cases the results show that better performance is achieved by allowing the decision to be deferred. The two model-parameters showing the best results were the absolute-residency time parameter and c0. They resulted in 645 and 768 errors, respectively. This can be explained by the observation that these particular features tend to capture the differences in confused words more so than the other features used. These features correlate well with what is seen when studying a spectrogram. The overall energy and duration of particular events are very important for word recognition. These two features bring out these characteristics. For every model-parameter used, an increase in error rate was found when the option of deferring to the baseline score was eliminated. For the two parameters mentioned above, the errors produced were 802 and 918, respectively, when deferring was not allowed. This was a bit surprising considering only a single model-parameter was used for any pass through the data base.

Experiment 3 — Fixed Variance

Since the results using just a single model-parameter showed improved performance over the complete model-parameter set, the suspicion that statistical noise might be overwhelming the recognizer became a genuine concern. This was thought to be a result of limited training data, resulting in small sample sets used for estimation. Poor estimates of this sort can alter the decision boundaries of a discriminator. This can result in degraded performance by the system. The first solution to this attempted to repeat experiment 2 without using variance in the discriminations. Since only a single parameter was to be used for any given pass through the data base, underlying variance differences between model-parameters did not present a problem. This experiment was aided by the discriminant calculations requiring no logarithms to be calculated. This enabled the experiment to run quickly. Results showed that the suspicions, mentioned above, were correct. These results are presented in Figure 5-2. The solid squares show points where deferring was used, and the diamonds show points where deferring was not used. Once again, the conclusion is that it is better to be able to defer the decision. Improved performance over the baseline was found with several of the parameters. The errors for the absolute-residency time parameter and c0 dropped to 582 and 621 errors, respectively. This represents final error rates of 6.4 and 6.8 percent. This is a significant improvement over the previous experiment.

Experiment 4 — Limited Model-Parameter Discrimination

The results of experiment 3 showed improvements when c0 and absolute duration were separately used as discriminators. Since these are quite different parameters, it was suspected that the improvements seen from each of these were somewhat orthogonal. If this were the case, even further improvements could be seen by combining these parameters in the scoring scheme. Two approaches were investigated. For each word model, two scores resulted. These correspond to c0

and absolute duration. The first approach used the decision from the second stage only if both of these scores pointed to the same word. In the second approach if no clear decision resulted from a parameter and the other parameter showed a clear decision, then the second-stage choice would be used. Both of these approaches showed improvement. Figure 5-2 shows the complete results of this experiment. The first scheme (both) resulted in 571 errors, and the second (eith) resulted in 539 errors. These are error rates of 6.3 and 5.9 percent. Table 5-5 gives the tabulated results of experiments 2, 3, and 4.

Experiment 5 — Multimodel-Parameters with Fixed Variance

Since combining parameters into the scoring procedure in a limited capacity showed some improvement in experiment 4, the use of all model-parameters was an approach that resurfaced. The use of these parameters was hoped to show significant improvement with the elimination of the problems associated with the poor variance estimates.

To get around poor variance estimation, grand variance estimates were made for each of the model-parameters except the duration models. These estimates were generated from all training tokens for all words for a given speaker. Using these variance estimates has the effect of weighting the contributions of each of the model-parameters, as described in Section 4. This provided the single most significant improvement of all experiments conducted. This system produced a total of 420 errors. This is a 4.6-percent error rate.

Experiment 6 — The Statistical T-Test

This experiment was motivated by two ideas. The first was that instances of model-parameters in particular nodes could prove to contribute no significant information to the final discrimination score. The second idea was that a system that could eliminate many parameters from the discrimination could save computation time as well as making the inclusion of confusion-specific features much simpler. The T-statistic threshold set at a significance level of 0.05 was found to eliminate more than 50 percent of all model-parameter discrimination instances from the scoring.

The results of this run showed a slight improvement over the same system without the T-test. The total number of errors was 413 as compared to 420 without the T-test. The significance of this result is that almost half the parameters used for discrimination were discarded without degrading performance. The slight increase in performance suggests that using only parameters that are clearly distinct is a more robust scheme than using many parameters that might be only contributing noise to the final score.

Experiment 7 — An Added Vector of Features

To test the ability of the discriminant system to handle many new features, an experiment was performed using an additional vector of seventeen, per-node features. The features were chosen to make use of longer-term spectral changes in the speech signal. The mean value of all

TABLE 5-5
Detailed Analysis of Experiments 2 and 3

Parameter	Estimated Variance		Fixed Variance	
	Defer	No Defer	Defer	No Defer
rel duration	1070	1251	832	1377
abs duration	645	802	582	667
c0	768	918	621	715
c1	872	996	675	759
c2	877	1054	694	783
c3	859	1017	722	818
c4	957	1132	803	877
c5	854	1015	714	845
c6	906	1066	745	811
c7	979	1150	696	807
c8	886	1048	798	922
c9	891	1087	779	893
c10	907	1110	796	910
c11	894	1083	848	1010
c12	1004	1186	917	1096
c13	966	1135	856	1015
c14	940	1147	838	1010
c15	930	1117	843	974
c16	970	1188	802	938
both	642		571	
either	632		539	

All numbers correspond to errors out of 9135 test tokens. The first column indicates which parameter was used in the discriminations. The second and third columns used variance estimates for each occurrence of the discriminating parameter. The fourth and fifth columns used fixed variance estimates for each of the parameters. Schemes where the score could be deferred and could not be deferred were both used.

cepstral parameters was calculated for each node. The difference in these mean values from adjacent nodes was used as the feature vector. Experiments 4 through 6 used 18 discriminant parameters. These parameters were c0 through c16 and a duration model. Seventeen more parameters were added in this experiment. A total of 35 parameters were thus used in this experiment. T-testing was included in the discriminations.

The result showed a final error rate of 4.4 percent, totaling 406 errors. This was slightly less than the 413 errors in experiment 7.

Experiment 8 — Threshold Modification

At this point it was observed that the second-stage system proved to be correct 98 percent of the time it made the final decision. Because of this, it was felt that a net improvement might be seen if more words were initially passed to the second stage. This could be realized by widening the threshold that decides when a word goes to the second stage. The threshold was reset at 1.0, and the identical experiment was performed as experiment 7. The result again showed improvement, but no drastic improvement. The total errors were now 390 or 4.3 percent. It is important that there was no significant change in error rate for this experiment. A hard threshold exists in the system that must be estimated. Ideally, the performance of a system should not be sensitive to variations of such a threshold. With the limited experimentation done in this area, the system displays an insensitivity to variations in the discriminating threshold.

Experiment 9 — Three-Way Discrimination

Results from experiment 8 demonstrated that many of the remaining errors were attributable to the correct word not being among the baseline's top two candidates. More than two-thirds of all errors from experiment 8 were a result of this situation. This experiment looked at the top three candidates from the baseline system when discriminations were required. A total of three scores was now computed corresponding to each pairwise combination of the three words being discriminated. The decision scheme used was to accept the second-stage score only, if both comparisons for any one word pointed to that word as the selection. Each pairwise decision was of the type found in Table 4-2.

The error rate decreased substantially. The error total was 351 or 3.8 percent. The percent of these errors, attributed to the correct word not being in the top three candidates from the baseline, was now only 41 percent.

Experiment 10 — A Modified Decision Scheme

Even though experiment 9 produced significant improvements, a large fraction of all errors was a result of the second stage deferring back to the baseline score. Over 40 percent of these deferred words produced errors. To reduce this effect, a scheme was sought that would greatly reduce the frequency of the second stage deferring back to the baseline.

This experiment differed from the previous experiment only by the pairwise decision scheme used. In this experiment the pairwise scheme was of the form found in Table 4-1, instead of the scheme found in Table 4-2.

The results of this experiment supported the ideas motivating it. The error total fell to 325, as compared to 351 in experiment 9, with a much smaller percent being attributable to deferrals. The error rate of this experiment was 3.5 percent.

Experiment 11 — Baseline Parameter Comparison

The following experiments were designed primarily to investigate various questions that arose during the period that the research was being conducted. Often an improvement in net performance was not expected from these experiments. Generally, the experiments try to isolate a particular feature of the system to evaluate its validity as a part of the larger system.

After obtaining many of the above-mentioned results, the question of why improvements were occurring was considered. If the reason was completely attributed to the added features such as duration, a better solution might just be to add these features into the baseline system. If the improvements came from new information that was obtained during the training procedure that the baseline does not use, the system would prove to be a significant improvement.

Experiment 11 involved a system similar to the systems described in the last few experiments. The major difference was that only those parameters available to the baseline system were used as discriminating parameters. These were c1 through c16. No duration model was used. The system used fixed variance, selective deferring to the baseline, and T-testing. The final error rate was 4.8 percent or 436 errors out of 9135 tokens. This was exactly the result that was hoped for. The improvement, compared to the 7.7-percent error rate of the baseline, suggests that more information is being obtained from these parameters than was used with the baseline. The poorer performance, as compared to other systems described, reflects the power of the system in its ability to incorporate added features into the scoring scheme.

Experiment 12 — An Alternative to T-Testing

In an experiment conducted by Rabiner and Wilpon,[22] a measurement that was in some ways similar to a T-statistic was applied to weight the effects of various parameters used in a discrimination. The sample size used to obtain the estimates was not taken into account.

$$W = \frac{|\hat{X} - \hat{Y}|}{\sqrt{S_x^2 + S_y^2}} \tag{5.1}$$

Equation (5.1) defines the weighting scheme. The log-likelihood ratio from a particular parameter is weighting by this value. With the statistical T-test this weight is either 0 or 1, depending on where the threshold lies. The T-test makes a binary decision rather than weighting. Rabiner and Wilpon's weighting scheme was installed on the current system and tested with the

system used in experiment 6, that resulted in a final error rate of 4.5 percent. Rabiner and Wilpon's scheme resulted in a total of 531 errors or 5.8 percent and, thus, degraded performance. Another approach to the focus-of-attention problem in discrimination is presented by Moore in Reference 23.

Experiment 13 — Effects of T-Testing with Feature Vectors

All experiments conducted after the initial T-testing run (experiment 6) included the use of T-testing in the discriminant system. An experiment was performed to measure how significant the T-testing was for systems using many added features. The experiment conducted was identical to experiment 8 with the T-testing mechanism removed. The final error rate was 4.4 percent or 398 errors. This compares to the same system with the T-test that resulted in 390 errors. The conclusion to be drawn from this is the following: the use of fixed variance compensates well for the originally poor estimates used in experiment 1. With the use of fixed variance, the likelihood ratio is a more reliable statistic to use in discrimination.

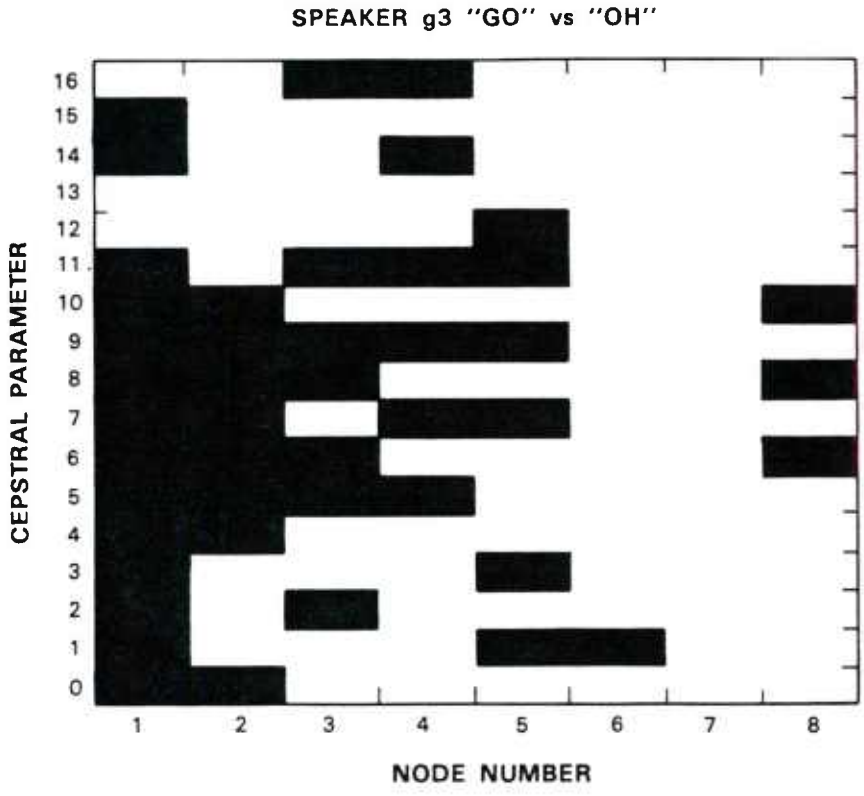
This is brought out by the observation that removing nearly half the discriminators with the T-test does not significantly improve performance. The parameters sifted out, therefore, do not significantly degrade performance. The T-test is still a powerful tool. The mechanism for removing parameters that may degrade performance is an important feature to a recognizer even though this facet was not seen in this case. The advantage of removing half of all computations for the second-stage system is obvious for any real-time application.

Experiment 14 — The Correlated Observation Assumption

This experiment was intended to demonstrate that the assumption of uncorrelated observation frames degrades recognizer performance. The system used was identical to that of experiment 10 with one exception, the assumption was made that observation frames were uncorrelated. The normalization, which formerly was done in the discriminant scoring to give nodes rather than observations equal weight, was eliminated. The error rate rose from 3.5 to 4.4 percent, showing that the assumption of independent observations is a poor one.

Focus of Attention

The T-test mechanism selectively discards parameters that do not contribute well to the discrimination. The parameters that are included should correspond to the parameters that intuitively focus the attention of the discrimination to key parts of the words. The key parts should correspond to regions where a great discrepancy is seen between the two words. For certain confusions this idea was examined more closely. Statistics were kept defining those parameters in these confusions that were kept or discarded during discrimination. Figure 5-3 demonstrates that most of the parameters used in a discrimination between "go" and "oh" were focused toward the beginning of the word. The vertical axis plots the parameters that were or were not used in the discrimination. The horizontal axis plots node number, a value that is monotonic with time. The



78054-4

Figure 5-3. Cepstral parameters used to discriminate the word models for "go" and "oh" are indicated with darkened regions. Most of the parameters used are concentrated toward the beginning nodes.

dark rectangles in the figure show instances where both word models used a particular parameter at a particular node. Figure 5-4 demonstrates similar information from an "eight/eighty" comparison. These plots agree with the intuition toward such discriminations. Words whose only acoustic difference is found at the beginning or end of the word will best be discriminated by focusing on that part of the word.

Figures 5-3 and -4 show parameters used by both word models for each node. The "go/oh" comparison is focused toward the beginning of the word, and the "eight/eighty" comparison is focused toward the end of the word.

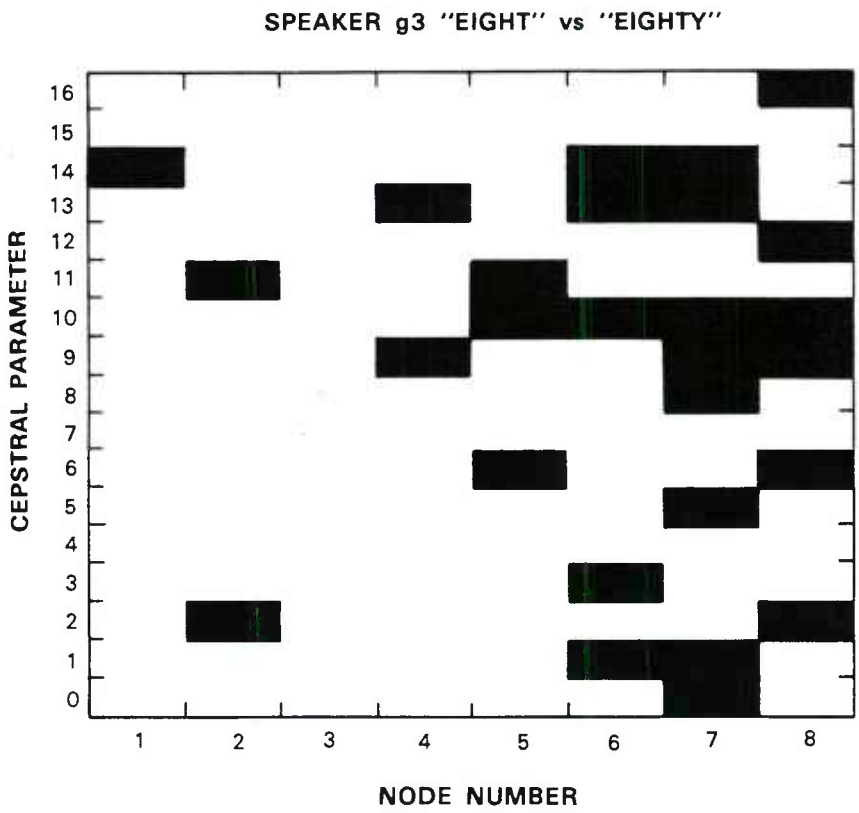


Figure 5-4. Similar to Figure 5-3 for the "eight" and "eighty" discrimination. Most of the parameters used are concentrated toward the end nodes.

6. SUGGESTIONS FOR FUTURE RESEARCH

This section describes some proposed modifications and extensions to the two-stage discriminant system.

6.1 WAVEFORM FEATURES

The discriminant system has demonstrated an ability to incorporate added features as discriminating parameters without degrading overall performance. In fact, the best system tested, experiment 10, included a total of 35 parameters used for discrimination.

Including features that may only address a limited subset of overall confusions is a possible technique for achieving improved system performance. The hope of adding such features is that they always will be available for inclusion in the discriminant scoring. Individual features only would be included though for resolving confusions where those particular features were significant. For confusions where the features bear no real significance, the T-testing mechanism should effectively sift out those features.

Possibilities for such features might include a nasality detector[24] or waveform-based features, as described in Reference 25. The strength of this system lies in its ability to incorporate many features without any concern over how often it might prove useful or if the adding of another feature might lessen the significance of another parameter.

6.2 SIFTING

A method which removed weak discriminating statistics from a classifier was proposed by Lippmann.[10] The T-testing algorithm described in some of the experiments is a primitive version of this idea.

For the system demonstrating the best overall performance, mean estimates were sifted out of the computation by the T-test, and estimated variances were replaced by grand variances. An extension of this idea is to make comparisons with the estimated variances similar to the comparisons of the estimated means. An F-ratio test can be used to make such comparisons. The threshold in this case is related to the ratio of the estimated variances. If the estimated variances are thought to be the same, the grand variance could be used in place of these estimates. Given enough training data, a similar test could be applied to covariance estimates between observation frames. The importance of this implementation is that the estimated moments that are thought to be significant are included in the classifier, and moments that are thought to be statistically insignificant are omitted.

6.3 ADAPTATION

The idea of having a recognition system that is constantly adapting itself by updating its models is a feature found to be difficult to incorporate in many systems. The complexity of the models should correspond to the amount of training data available.

When templates are first formed, quite often a small amount of training data are used. Using estimates of higher-order moments will probably degrade performance. After being used for some time an adaptive system will accumulate abundant training data, and higher-order estimates may then prove to be useful. Having adaptive models in a simple recognition system that might only use estimates of first-order moments will not completely exploit the increase in available data. Such a system should adapt into a more complex system by not only updating existing estimates, but by including higher-order moments when enough data are available.

The idea of sifting is an excellent vehicle for implementing such a system. When the training data are limited, perhaps only the means will pass the statistical significance tests (T-test, F-ratio test). As the amount of data increases the higher-order estimates will become more significant and will then be gradually incorporated into the scoring.

Using test data to augment word model estimates will enable the system to continually update and adapt itself. By keeping track of these new data the system will automatically use higher-order moments of various estimated statistics whose sample sets may have been too small to be considered significant prior to testing. Such a system thus automatically adapts its complexity to the amount of data available.

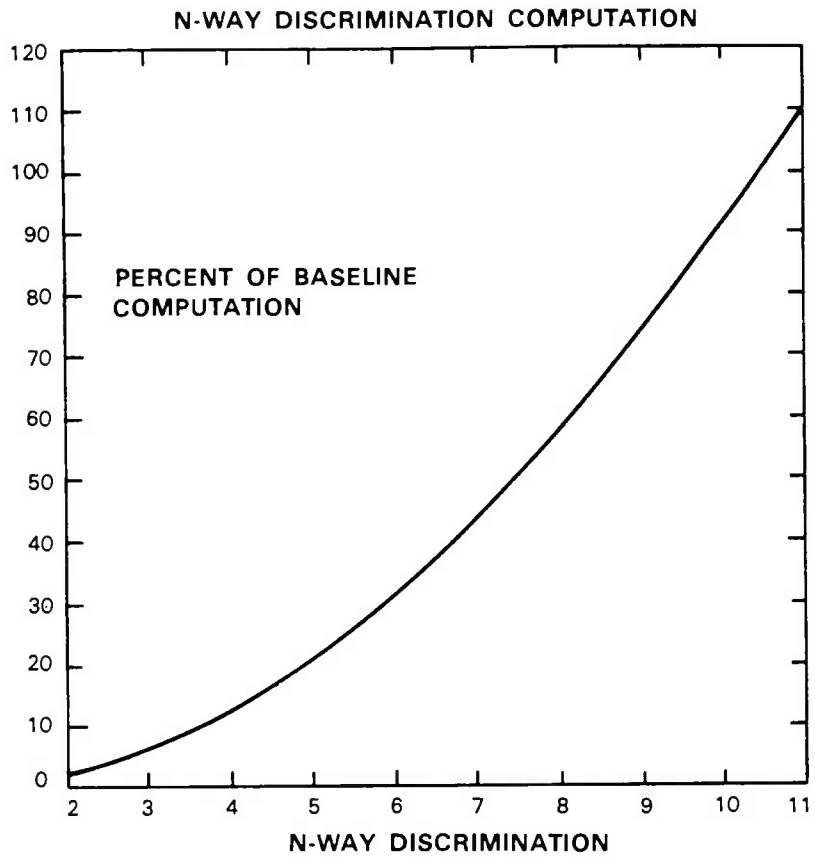
6.4 EXTENDED N-WAY DISCRIMINATION

Experiments 9 and 10 demonstrated a significant improvement in overall system performance by looking at more than just the top two candidates from the baseline system. Further improvements might be seen by extending this idea to a general N-way discrimination.

The most efficient way to handle this approach is to have the depth of the search be dependent on the degree to which the baseline scores were clustered. If just the top two scores are very close and the third is distant from these two, a two-way discrimination should be adequate. If the top four scores are all clustered together, a four-way discrimination might be needed.

The only real limitation on such an implementation is computation time. The discriminant system used for the experiments described in Section 5 required about two percent of the computation time required by the baseline system. This two percent corresponds to a two-way discrimination. As the number of candidates N used in an N-way discrimination increases, the computation required increases by a factor of $N!/2(N-2)!$. Figure 6-1 shows how the computation time of the discriminator rises as a function of the number of candidates used. Surprisingly, as many as ten candidates can be discriminated before the computation time of the discrimination equals that of the baseline.

There are many ways in which an N-way discrimination problem could be collapsed into an identification problem. N-way discrimination was used in this study because it allows for a much simpler implementation of sifting. Further research could examine identification procedures that include sifting.



78054-3

Figure 6-1. Vertical axis shows the percent of the baseline computation that is necessary for an N-way discrimination. The horizontal axis plots N.

6.5 USE WITH LARGER SYSTEMS

The greatest burden the discriminant system described thus far places on an implementation is memory storage. For the 35-word vocabulary, statistics must be kept for discriminating between 595 different word pairs.

An idea proposed earlier was to divide the full vocabulary into several subvocabularies. Statistics then would be kept only on word pairs of words in common subvocabularies. The thought behind this approach is that the first stage will usually confuse words in common subvocabularies. This is probably true to a certain extent; however, some common confusions will occur across various subvocabularies.

A better solution would be to investigate nearest-neighbor scores during the training phase when all training tokens are recognized by the baseline. By only saving statistics on word pairs that appear as the top N candidates for any token of this training phase, the amount of storage can be greatly reduced.

Using two-way discrimination as the recognition scheme, and the statistic-saving scheme described above, only 10.4 percent of all word pairs were saved. This represents the statistics saved from the top two candidates. During recognition, discriminations were possible with the trimmed data 83.9 percent of the time a discrimination was called for.

Using a statistic-saving scheme where all word pairs from the first three candidates during training are saved, the storage requirement is 23.7 percent of all possible word pairs. During recognition with these statistics saved, 94.5 percent of all attempted discriminations were possible.

For the baseline HMM system with a 35-word vocabulary, 45 kbytes of storage is needed for statistics on the 16 cepstral parameters. For the 16 cepstral parameters and c0, the discriminant system requires 1.3 Mbytes of storage. Using grand variances instead of the variance estimates halves this storage requirement. The use of T-testing drops the storage even further to 333 kbytes. If the nearest-neighbor schemes for generating statistics are used, the requirement falls to 79 kbytes for the top three candidate statistics, and 35 kbytes for the top two candidate statistics.

7. CONCLUSIONS

The major conclusion to be drawn is that the two-stage discriminant system works well. The error rate was more than halved, resulting in a drop from 7.7 to 3.5 percent. Many references were made to other work which was done investigating ideas similar to those presented here. Four key differences can be found between the work presented in this report and the work referred to in other papers. The first is that the weighting scheme used in Rabiner and Wilpon[26] was shown to be less effective than the binary scheme used in experiment 12. The number of samples used in the estimates were not taken into account by Rabiner and Wilpon, and a linear instead of binary weight was used. The second key difference is that the effects of limited training data were explicitly considered in this report. This effect has not been addressed previously in the manner described in this report. Fixed variance and T-testing were added to address this problem. This also allowed new feature parameters to be added without degrading performance. The third key difference is that observations were treated as correlated within nodes. Experiment 14 demonstrated that treating observations as uncorrelated degrades performance. The final difference is that the duration model used modeled node duration explicitly. This was based on the actual number of observations assigned to a node. Experiments 2, 3, and 4 demonstrated that this model performed better than the duration model based on relative duration used in Reference 27.

The discriminant system, as it now stands, is a very good tool for performing experiments. Many features can be quickly and easily added without any structural change to the system. This was demonstrated by experiment 7 which doubled the number of discriminant parameters.

Finally, the discriminant system can be reasonably implemented for real applications. The storage requirement for a vocabulary the size used in this work can be as low as 35 kbytes. The computation requirement is also less than half that of the baseline for discriminations using as many as the top seven candidates.

REFERENCES

1. E. Bocchieri and G.R. Doddington, "Frame Specific Statistical Features for Speaker Independent Speech Recognition," submitted to IEEE Trans. ASSP, September 1985.
2. G.R. Doddington and T.B. Schalk, "Speech Recognition: Turning Theory to Practice," IEEE Spectrum (September 1981), pp. 26-31.
3. J.K. Baker, "Stochastic Modeling for Automatic Speech Understanding," in *Speech Recognition*, edited by R. Reddy (Academic Press, New York, 1975).
4. F. Jelinek, "The Development of an Experimental Discrete Dictation Recognizer," Proc. IEEE **73**, No. 11, 1616-1624 (November 1985).
5. D.B. Paul, R.P. Lippmann, Y. Chen, and C.J. Weinstein, "Robust HMM-Based Techniques for Recognition of Speech Produced Under Stress and Noise," Speech Tech 86 Conf. Rec., Media Dimensions, New York, April 1986.
6. A.B. Poritz, "Linear Predictive Hidden Markov Models and the Speech Signal," IEEE ICASSP '82 Proc., April 1982.
7. L.R. Rabiner, B.H. Juang, S.E. Levinson, and M.M. Sondhi, "Recognition of Isolated Digits Using Hidden Markov Models With Continuous Mixture Densities," AT&T Tech. J. **64**, No. 6, 1211-1234 (1985).
8. R.P. Lippmann, M.A. Mack, and D.B. Paul, "Multi-Style Training for Robust Speech Recognition in Stress," presented at the ASSP Conference, May 1986.
9. R. Duda and P. Hart, *Pattern Classification and Scene Analysis* (John Wiley and Sons, New York, 1973).
10. R.P. Lippmann, personal communication, 1986.
11. P.K. Rajasekaran, G.R. Doddington, and J.W. Picone, "Recognition of Speech Under Stress and in Noise," IEEE ICASSP '86 Proc., April 1986.
12. S.E. Levinson, L.R. Rabiner, and M.M. Sondhi, "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition," Bell Syst. Tech. J. **62**, No. 4, 1035-1074 (1983).
13. L.A. Liporace, "Maximum Likelihood Estimation for Multivariate Observations of Markov Sources," IEEE Trans. Inf. Theory **IT-28**, No. 5 (September 1982).
14. H.R. Jex, "A Proposed Set of Standardized Sub-Critical Tasks for Tracing Workload Calibration," in N. Moray, *Mental Workload: Its Theory and Measurement* (Plenum Press, New York, 1979), pp. 170-188.

15. E. Lombard, "Le Signe de l'Elevation de la Voix," Ann. Maladiers Orielle, Larynx, Nez, Pharynx **37**, 101-119 (1911).

16. V.W. Zue *et al.*, "The Development of the MIT Lisp-Machine Based Speech Research Workstation," IEEE ICASSP '86 Proc., April 1986.

17. J.K. Baker, J.M. Baker, R. Roth, and P. Bamberg, "Cost Effective Speech Processing, IEEE ICASSP '84 Proc., 19-21 March 1984.

18. P. Hoel, *Introduction to Mathematical Statistics* (John Wiley and Sons, New York, 1984).

19. D.V. Huntsberger and P. Billingsley, *Elements of Statistical Interference* (Allyn and Bacon, Inc., Boston, 1981).

20. F. Jelinek, R.L. Mercer, and L.R. Bahl, "Continuous Speech Recognition: Statistical Methods," in *Handbook of Statistics*, edited by P.R. Krishnaiah and L.N. Kanal (Elsevier North-Holland, New York, 1982), Vol. 2, pp. 549-593.

21. G.W. Snedecor and W.G. Cochran, *Statistical Methods* (The Iowa State University Press, 1980).

22. L.R. Rabiner and J.G. Wilpon, "A Two-Pass Pattern-Recognition Approach to Isolated Word Recognition," Bell Syst. Tech. J. **50**, No. 5, 739-766 (May 1981).

23. R.K. Moore, M.J. Russell, and M.J. Tomlinson, "The Discriminative Network: A Mechanism for Focusing Recognition in Whole-Word Pattern Matching," IEEE ICASSP '83, Proc., Boston, 14-26 April 1983.

24. J.R. Glass, "Nasal Consonants and Nasalized Vowels: An Acoustic Study and Recognition Experiment," SM Thesis, Dept. of Electrical Engineering and Computer Science, MIT, December 1984.

25. J.M. Baker, "A New Time-Domain Analysis of Human Speech and Other Complex Waveforms," PhD Thesis, Dept. of Computer Science, Carnegie-Mellon University, 1975.

26. D. Pisoni, R.H. Bernacki, H.C. Nusbaum, and M. Yuchtman, "Some Acoustic-Phonetic Correlates of Speech Produced in Noise," IEEE ICASSP '85 Proc., 26-29 March 1985.

27. L.R. Rabiner, S.E. Levinson, and M.M. Sondhi, "On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent, Isolated Word Recognition," Bell Syst. Tech. J. **62**, No. 4, 1075-1104 (1983).

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) TR-773	5. MONITORING ORGANIZATION REPORT NUMBER(S) ESD-TR-87-026
6a. NAME OF PERFORMING ORGANIZATION Lincoln Laboratory, MIT 6b. OFFICE SYMBOL *(If applicable)*	7a. NAME OF MONITORING ORGANIZATION Electronic Systems Division
6c. ADDRESS *(City, State, and Zip Code)* P.O. Box 73 Lexington, MA 02173-0073	7b. ADDRESS *(City, State, and Zip Code)* Hanscom AFB, MA 01731
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Defense Advanced Research Projects Agency 8b. OFFICE SYMBOL *(If applicable)* DARPA/ISTO	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER Program 337 F19628-85-C-0002
8c. ADDRESS *(City, State, and Zip Code)* 1400 Wilson Boulevard Arlington, VA 22209	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 62301E; PROJECT NO.; TASK NO.; WORK UNIT ACCESSION NO. AO 5328

11. TITLE *(Include Security Classification)*

A Two-Stage Isolated-Word Recognition System Using Discriminant Analysis

12. PERSONAL AUTHOR(S)
Edward A. Martin

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT *(Year, Month, Day)*	15. PAGE COUNT
Technical Report	FROM — TO —	5 August 1987	62

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18 SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*		
FIELD	GROUP	SUB-GROUP	robust speech recognition	discriminant analysis	two-stage recognition
			HMM word recognition	speech recognition	limited training data
			Hidden Markov Model	simulated stress	

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

This report describes a two-stage isolated-word recognition system using a Hidden Markov Model (HMM) recognizer in the first stage, and a statistical discriminator in the second stage. The second-stage system performs pairwise discriminations between the top few candidate word models when no clear decision is made from the first stage. Likelihood-ratio comparisons and a new technique called "sifting" are used to focus attention on those features that best differentiate word pairs.

This system alleviates four fundamental problems which are found with most conventional speech recognition systems. These problems include: (1) the effects of limited training data are not explicitly taken into account, (2) the correlation between adjacent observation frames is incorrectly modeled, (3) durations of acoustic events are poorly modeled, and (4) features which might be important in discriminating only among specific word pairs or sets of words are not easily incorporated into the system without degrading overall performance. The system was tested on a 35 word/10,000 token stressed-speech isolated-word data base created at Lincoln Laboratory. The adding of the second-stage discriminating system reduced the error rate by more than a factor of 2. The overall error rate fell from 7.7 percent with only the HMM system to 3.5 percent with both the HMM system and the discriminator.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION	
☒ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS	Unclassified	
22a. NAME OF RESPONSIBLE INDIVIDUAL Maj. Thomas J. Alpert, USAF	22b. TELEPHONE *(Include Area Code)* (617) 863-5500, Ext. 2330	22c. OFFICE SYMBOL ESD/TML